Journal of Informetrics 8 (2014) 478–485

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Empirical study of *L*-Sequence: The basic *h*-index sequence for cumulative publications with consideration of the yearly citation performance

Yu Liu [a,*], Yongliang Yang [b]

[a] School of Software, Dalian University of Technology, Dalian 116024, PR China
[b] Center for Molecular Medicine, School of Life Science and Biotechnology, Dalian University of Technology, Dalian 116024, PR China

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form 14 February 2014
Accepted 19 March 2014
Available online 19 April 2014

Keywords:
Bibliometrics
Citations
Research evaluation
H-index sequence

ABSTRACT

Most current *h*-type indicators use only a single number to measure a scientist's productivity and impact of his/her published works. Although a single number is simple to calculate, it fails to outline his/her academic performance varying with time. We empirically study the basic *h*-index sequence for cumulative publications with consideration of the yearly citation performance (for convenience, referred as *L*-Sequence). *L*-Sequence consists of a series of *L* factors. Based on the citations received in the corresponding individual year, every factor along a scientist's career span is calculated by using the h index formula. Thus *L*-Sequence shows the scientist's dynamic research trajectory and provides insight into his/her scientific performance at different periods. Furthermore, L_{α}, summing up all factors of *L*-Sequence, is for the evaluation of the whole research career as alternative to other *h*-index variants. Importantly, the partial factors of the *L*-Sequence can be adapted for different evaluation tasks. Moreover, *L*-Sequence could be used to highlight outstanding scientists in a specific period whose research interests can be used to study the history and trends of a specific discipline.

1. Introduction

Bibliometrics has played an increasingly important role in evaluating individual researchers. It could be used as a quantitative analysis and assessment tool in tasks such as faculty promotion, funding allocation and awarding scientific prizes, etc (King, 1987). Bibliometrics provides a reliable and cost–effective way to evaluate scientific publications and their citations compared to the resource-expensive peer review (Abramo & DAngelo, 2011).

In 2005, a simple indicator for the assessment of the academic performance was suggested by Hirsch (2005), with consideration of both productivity and impact. The *h*-index has received a lot of attention from the scientific community in the last few years owing to its excellent properties (Ball, 2005). Many variants of the *h*-index were proposed to improve the original *h*-index (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Egghe, 2010; Zhang, Thijs, & Glänzel, 2011). However, most of these *h*-type indicators use only a single number to measure scientists' life-long performance. One dimensional indicator lacks the ability to reveal the evolution details of a scientist's career at different periods.

* Corresponding author. Tel.: +86 15524582688.
E-mail address: yuliu@dlut.edu.cn (Y. Liu).

http://dx.doi.org/10.1016/j.joi.2014.03.002
1751-1577/

In contrast to single number, a series of index can describe a scientist's academic performance along with his/her career period. Liang first proposed an *h*-index sequence that was calculated in the reverse direction (Liang, 2006). However, Egghe pointed out that the calculation in the forward direction is more practical and easy to understand (Egghe, 2009b). Another interesting work is by Liu and Rousseau (2008) in which they defined 10 types of *h*-index sequences. Unfortunately, empirical study of practical examples based on large datasets are still lacking. Further, Egghe investigated four important sequences of which three sequences were defined by Liu & Rousseau before. These sequences were well explained and some practical examples were also discussed. Egghe's study has stimulated more research in this direction: for example, Fred Y. Ye and Ronald Rousseau studied the relationship between the power law model and total career *h*-index sequences (Ye & Rousseau, 2008), denoted as h_4 in Egghe's sequences; Wu, Lozano and Helbing performed the empirical study of the real career *h*-index sequence (Wu, Lozano, & Helbing, 2011), denoted as h_3 in Egghe's work. Based on h_3, Lin Zhang and Wolfgang Glänzel proposed the age dynamics of its *h*-core. Moreover, the other two time series, evolution of co-authorship and the age pyramids, were also presented in order to capture various facets of individual academic careers (Zhang & Glänzel, 2012).

In this paper, we performed the empirical study of another important *h*-index sequence with consideration of yearly citation performance for cumulative publications, denoted as h_2 in Egghe's work. Here we present the sequence as *L*-Sequence for convenience. As stated by Egghe, the sequence is rather necessary and challenging to study. The fact of slightly increasing in the sequence should be further studied and interpreted (Egghe, 2009a). We use a large bibliographic data set for computer scientists to evaluate the performance of *L*-Sequence. Our experimental results demonstrate that *L*-Sequence could effectively reflect the dynamic properties of a scientist's productivity and citation impact. Particularly, L_∞, summing up all factors of *L*-Sequence, can be used for the evaluation of the whole research career of a scientist as an alternative to other *h*-index variants. Importantly, the partial factors of the *L*-Sequence can be adapted to various evaluation contexts. In addition, the scientific impact of researchers is normalized to the year coordination. Hence, it is easy to compare researchers' performance in specified years, which becomes feasible to study the history and trends of a research field or discipline.

2. *L*-Sequence

Instead of using a single number, *L*-Sequence uses a sequence of measuring factors along with a scientist's career period to reflect his/her research performance. A series of process variables over the time can provide an accurate and sufficient description of a scientist's dynamic research trajectory, consistent with underlying natural mechanism of scientific research.

2.1. Definition of L-Sequence

Suppose a scientist has published n papers, $P_1, P_2, \ldots, P_n$, along his/her research career. The year of the first publication is T_1 and the current year is T_2. *L*-factor of each year (L_t) is calculated by the h index formula based on the citations received in year t for all papers. Thus, a series of *L*-factor for each year $L_{T_1}, L_{T_1+1}, \ldots, L_{T_2}$ constitute the *L*-Sequence.

2.2. Graphical illustration of L-Sequence

To explicitly understand the *L*-Sequence, Fig. 1 demonstrates the calculation process based on real bibliographic data of Judea Pearl, 2011 Tuning Award Winner. The citation counts for each of his papers in each individual year are recorded and then plotted in a single color as in Fig. 1. The citation counts received in one year are defined as the citation slice for this year, which is subsequently used to calculate a factor of *L*-Sequence. For example, the citation slice in 2000 is employed to calculate L_{2000}. In the top left corner of Fig. 1, citation slice in 2000 is displayed in the form of paper-citation distribution and the *h*-index formula is used to calculate L_{2000} based on the citation counts received in 2000. In the same way, a series of factors which consists of *L*-Sequence, are achieved one year by one year during a scientist's career. Thus Judea Pearl' *L*-Sequence is shown in Fig. 2.

Notably, in the course of calculating *L*-Sequence, a highly-cited paper for a long period could be involved in the calculation of L_t factors for several years, thereby making substantial contribution to the entire *L*-Sequence. We provided herein an example of a highly-cited paper from 1990 to 2010 as displayed as blue-color dash line (asterisk) in Fig. 1. Conversely, lowly-cited paper will be ignored in the calculation of L_t factors and outdated paper will barely make contribution to L_t factors when few citation counts were received. Again, we provided another example of an outdated paper from 1978 to 1988 as displayed as purple-color dash line (asterisk) in Fig. 1. This paper received its highest citation in 1981 and contributed to the L_t factors in the following two years. However, since 1984, this paper barely contributes to the calculation of L_t factors because it has received very few citations. This is consistent with the evolution process of research activities.

2.3. Application of L-Sequence

L-Sequence is a flexible evaluation tool and provides more details for researcher's academic performance varying with time. More importantly, users could select continuous or discrete parts of the factors in *L*-Sequence for different evaluation purposes. For instance, for awarding purpose, the performance of individual's entire research career should be taken into account. Therefore, all factors of *L*-Sequence could be summed up to indicate the life-long performance; For promotion and project support, the performance of recent m years is more valuable. Consequently, the recent m year factors could be

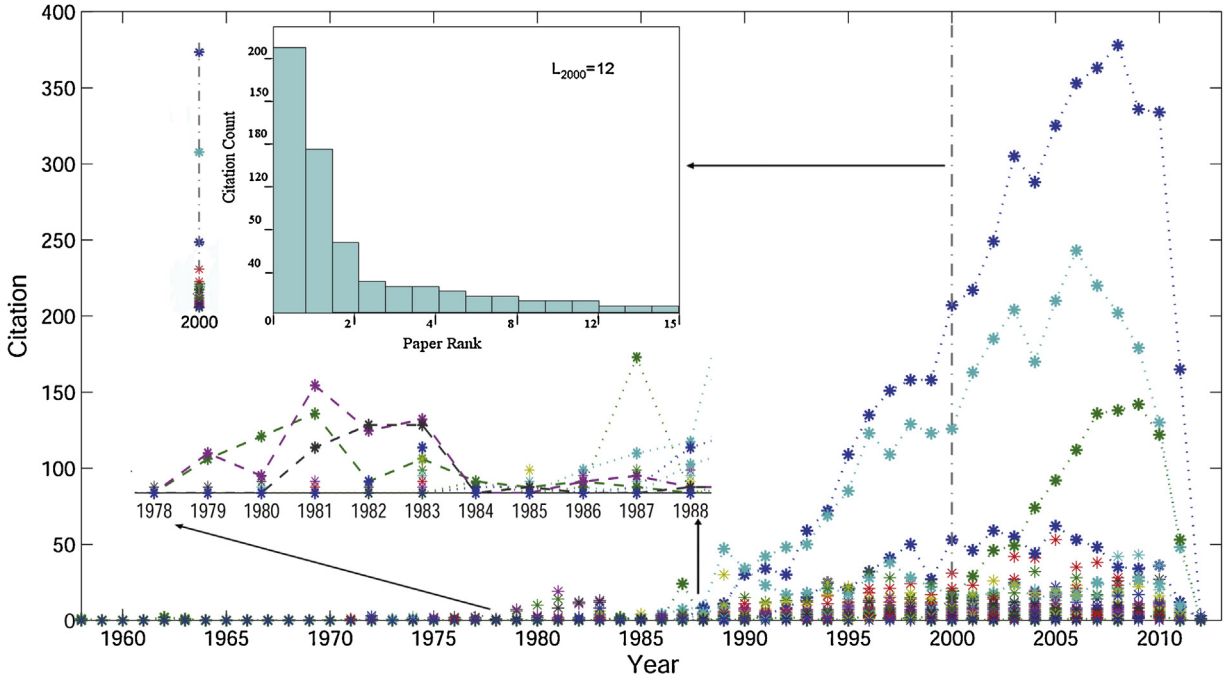


Fig. 1. The mechanism of *L* sequence calculation: Data of Judea Pearl, 2011 Tuning Award Winner's publications and corresponding citations. X-axis represents chronological years and Y-axis represents the citation counts. The asterisk represents the citation counts for a single paper displayed in different colors. *L*-factors are calculated based on a group of color-dots in one year and this group is called citation slice. The calculation of *L*-factor slice for 2000 is used as an example at the top left corner of the Figure. The bottom left of the Figure is used to address the situation that a paper does not always contribute to *L*-factors. This part was amplified from the data of 1976–1986. (For interpretation of the references to color in text near the reference citation, the reader is referred to the web version of this article.)

summed up; For discovering rising star, the increment of the recent *m* years with respect to the former *m* years could be used. Following are formulas mentioned above.

For whole career,

$$L_{\infty} = \sum_{t=T_1}^{T_2} L_t \tag{1}$$

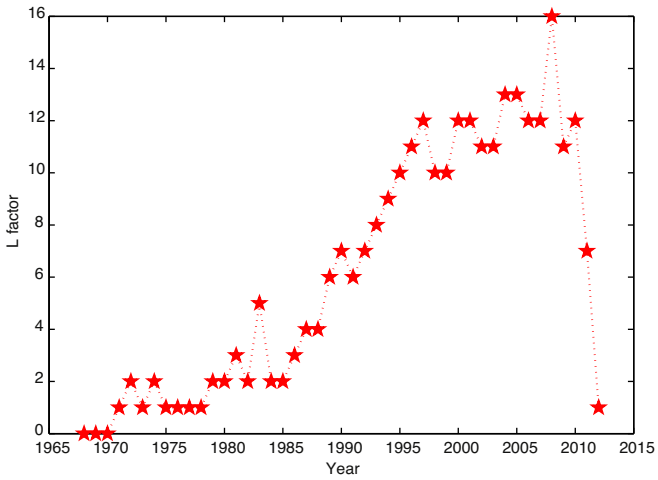


Fig. 2. The *L*-Sequence of Judea Pearl, 2011 Tuning Award Winner. X-axis represents chronological years and Y-axis represents the *L*-factor. The asterisk in red color represents the *L*-factor for each individual year. We want to remind the reader that the *L*-factors for 2011 and 2012 are calculated based on incomplete citation information, which will be released within future two years after publication.

Table 1
Comparison of researchers with $H = 42$

Name	Publications	Citations	H index	G index	L_α
Leslie Valiant	132	12,932	42	113	289
Robert Haralick	436	15,174	42	117	247
Stuart K. Card	155	11,842	42	108	237
Kenneth Birman	185	8956	42	93	223
Yehoshua Sagiv	206	7059	42	80	221
Kenneth L. McMillan	136	11,472	42	107	209
Steven K Feiner	244	9497	42	95	180
Kurt Mehlhorn	388	7798	42	75	178
Niraj K. Jha	461	6345	42	62	164
Rajkumar Buyya	432	8079	42	81	140

For year T_i to T_j

$$L_{T_i-T_j} = \sum_{t=T_i}^{T_j} L_t \tag{2}$$

For recent m years

$$L_{T_2 \bowtie m} = \sum_{t=T_2-m+1}^{T_2} L_t \tag{3}$$

For the increment of recent m years with respect to former m years

$$\Delta L_{T_2 \bowtie m} = \sum_{t=T_2-m+1}^{T_2} L_t - \sum_{t=T_2-2m+1}^{T_2-m} L_t \tag{4}$$

Moreover, *L*-Sequence has another valuable function. It can help us understand the history and trends of a specific discipline. This is because *L*-Sequence could align and compare a group of scientists in chronological sequence. Given specified years or periods, some representative scientists can be selected according to the values of the corresponding *L* factors. Then their research interests can be employed to study research history and trends.

3. Empirical study

In this section, we present three experiments which demonstrate the merits of *L*-Sequence as a flexible evaluation tool. All computer scientists' bibliographic data and citations were collected from the Microsoft Academy Website (http://academic.research.microsoft.com/). To this end, we designed a computer program to automatically extract citation counts, citing papers and publication time of the citing papers for scientists' publications so that the citation slice could be easily formed and the scientists could be filtered on a large scale. The top ranking scientists were listed in the results of corresponding experiments.

3.1. Comparison of L_α to h and g index

L_α, summing up all factors of *L*-Sequence, could be used for the evaluation of the whole research career of a scientist as an alternative to other *h*-type indicators. Here h and g index were compared with L_α. The citation data of Leslie G. Valiant, 2010 Turing Award winner, and other researchers with the same *h*-index or similar *g*-index to Leslie G. Valiant were used to validate the performance of L_α.

The Turing Award is recognized as the "highest distinction in Computer science" and "Nobel Prize of computing", awarded by the Association for Computing Machinery (ACM). Leslie G. Valiant was selected as the winner of the ACM Turing Award due to his extraordinary career in theoretical computer science, which means his impact is highly recognized by peer review. Leslie G. Valiant' *h*-index and *g*-index are 42 and 113 respectively.

First, Table 1 lists the bibliometrics data regarding to Leslie G. Valiant and other scientists who have the same *h*-index of 42. The identical *h*-index implicates that it fails to discriminate Leslie G. Valiant with other researchers. However, from the perspective of peer review, Leslie G. Valiant is supposed to be more influential in the field of computer science. This is mainly because *h*-index does not consider the excess citations received by papers within *h*-core. While *g*-index (Egghe, 2006a, 2006b) was able to eliminate the excess citation problem by weighting more on the highly-cited papers. Strikingly, it seems from Table 1 that L_α is well correlated with *g*-index by accounting for excess citations. Therefore L_α could also overcome the excess citation problem and then give a better discrimination among the scientists above. Although having some merits, the key drawback of *g*-index is that it could be disproportionately affected by a single very highly cited paper. A

Table 2
Comparison of researchers with $G \simeq 113$

Name	Publications	Citations	H index	G index	L_α
Leslie Valiant	132	12,932	42	113	289
William James Dally	281	13,394	50	113	259
Barry Boehm	471	14,124	47	113	253
James F. Kurose	380	14,477	59	113	245
William T. Freeman	238	13,445	52	114	236
Dieter Fox	220	13,942	60	116	229
Joseph Hellerstein	265	13,149	51	113	208
Pietro Perona	246	14,038	48	116	202
Gregor Kiczales	143	13,065	38	114	200
John Shawe-Taylor	315	13,244	38	113	161

typical example is shown in Table 1: Robert Haralick has the highest *g*-index of 117 among all scientists followed by Leslie G. Valiant with *g*-index of 113. The highest and the second highest citation for Robert Haralick are 3392 and 1570; the highest and the second highest citation for Leslie G. Valiant's are 2378 and 1439, respectively. It is the single publication of Robert Haralick with citation counts of 3392 responsible for his higher *g*-index. By contrast, the higher L_α of Leslie G. Valiant means that L_α could avoid to disproportionately affected by a single very highly cited paper.

Second, another group of scientists with *g*-index around 113 were listed in Table 2. The similar *g*-index scores implicate that *g*-index fails to discriminate these scientists. However, Leslie G. Valiant still obtains the highest score among them according to L_α.

In a word, in contrast with *h*-index and *g*-index, Leslie G. Valiant has the highest L_α among all the scientists. It seems that L_α has more discriminative power and the results ranking by L_α is analogous to the mechanism of peer review. Intriguingly, Leslie G. Valiant can be recognized as the best researcher in the above two groups by L_α in spite of some other researchers with more citations, higher *h*-index or *g*-index. This could be explained by the unique feature of *L*-Sequence that it changes over years and represents the performance of a scientist at different periods. Therefore, *L*-Sequence including process information (series of numbers can reflect persistent properties) is in accordance with the habits of human cognition.

In Fig. 3, the *L*-Sequences of five representative scientists are displayed over years. Among the five researchers, Leslie Valiant and Judea Pearl are recipients of Turing Award in 2010 and 2011, respectively. Leslie Valiant and Judea Pearl's *L* factors are continuously above 10 for more than 15 years. This accounts for the higher L_α of Leslie Valiant and Judea Pearl compared to other researchers. Therefore, for lifetime achievement evaluation, L_α strongly favors those who could maintain constant and high-profile scientific outputs over a long time.

3.2. Applicable to different evaluation objective

One additional merit of *L*-Sequence is that it could combine L_t factors from certain years for different evaluation tasks.

For example, $L_{2009-2011}$, the sum of factors in *L*-Sequence from 2009 to 2011, reflects the impact of active scientists within these three years. It could be used in the evaluation for job promotion or grant allocation.

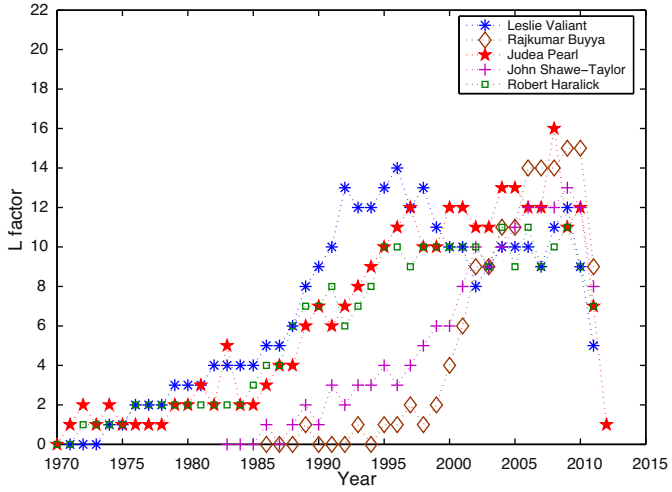


Fig. 3. Comparison of the *L*-Sequences for five representative scientists: Leslie Valiant, Judea Pearl, Robert Haralick, RajkumarBuyya and John Shawe-Taylor. Leslie Valiant and Judea Pearl are recipients of Turing Award 2010 and 2011. X-axis represents chronological years and Y-axis represents the *L*-factor. The different symbols represent the *L*-factor for each scientist.

Table 3
Ranking researchers with L_{α} and $L_{2009--2011}$

Rank	Name	L_{α}	Career length	Name	$L_{2009--2011}$	Career length
1	Jeffrey D. Ullman	531	51	David Donoho	76	31
2	Scott J. Shenker	431	31	Scott J. Shenker	70	30
3	Robert Endre Tarjan	426	41	David E. Culler	70	29
4	RTomaso A. Poggio	418	43	Michael I. Jordan	69	28
5	Richard Manning Karp	412	53	David Tse	69	44
6	Lotfi A. Zadeh	402	59	Andrew Zisserman	68	29
7	Donald E. Knuth	399	46	Deborah Estrin	66	27
8	Leslie Lamport	389	39	Anil K. Jain	69	28
9	Ronald L. Rivest	375	40	Jitendra Malik	61	31
10	Deborah Estrin	374	33	Stanley Oshe	69	45

Table 4
Ranking researchers with $L_{2009--2011} -- L_{2006--2008}$

Rank	Name	$L_{2009--2011} -- L_{2006--2008}$	Career length
1	Ying-chang Liang	26	19
2	Zidong Wang	22	19
3	Jeffrey G. Andrews	20	13
4	Jure Leskovec	19	9
5	Syed Ali Jafar	18	12
6	James Lam	17	30
7	Christian Kaestner	16	6
8	Yuri Bazilevs	16	7
9	Shuguang Cui	15	12
10	Ekram Hossain	15	12

$L_{2009--2011} -- L_{2006--2008}$ indicates the increment of scientist's achievement. This could help identify rising stars avoiding unfair comparison between junior researchers and scientists with a long career.

Table 3 lists top 10 L_{α} researchers on the left representing prominent scientists with life-long achievements whereas the top 10 $L_{2009-2011}$ researchers on the right representing prominent scientists who make a big impact in recent three years. Table 4 lists top 10 researchers as rising stars identified by $L_{2009--2011} -- L_{2006--2008}$ in computer science. Among these researchers, some have a career length about 20 years while others have only about 10 years of career length. From this perspective, *L*-Sequence is able to identify rising junior researchers with short career length.

Notably, one year factor might also be useful in academic evaluation and comparison. The box plots of one year factor of *L*-Sequence are shown in Fig. 4. We noticed that one year factor of top-notch scientists ranked by *L*-Sequence is usually larger than 12. Therefore, this may denote a reference value of outstanding scientists.

3.3. Investigation of developing trend for research discipline

As previously mentioned, *L*-Sequence could help to assess scientific impact in a specific time period. As *L*-Sequence involves a time coordinate, scientists could be compared with each other regardless of their age and career length.

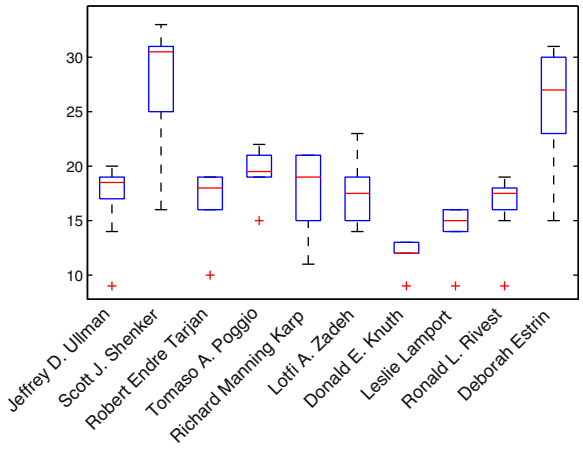
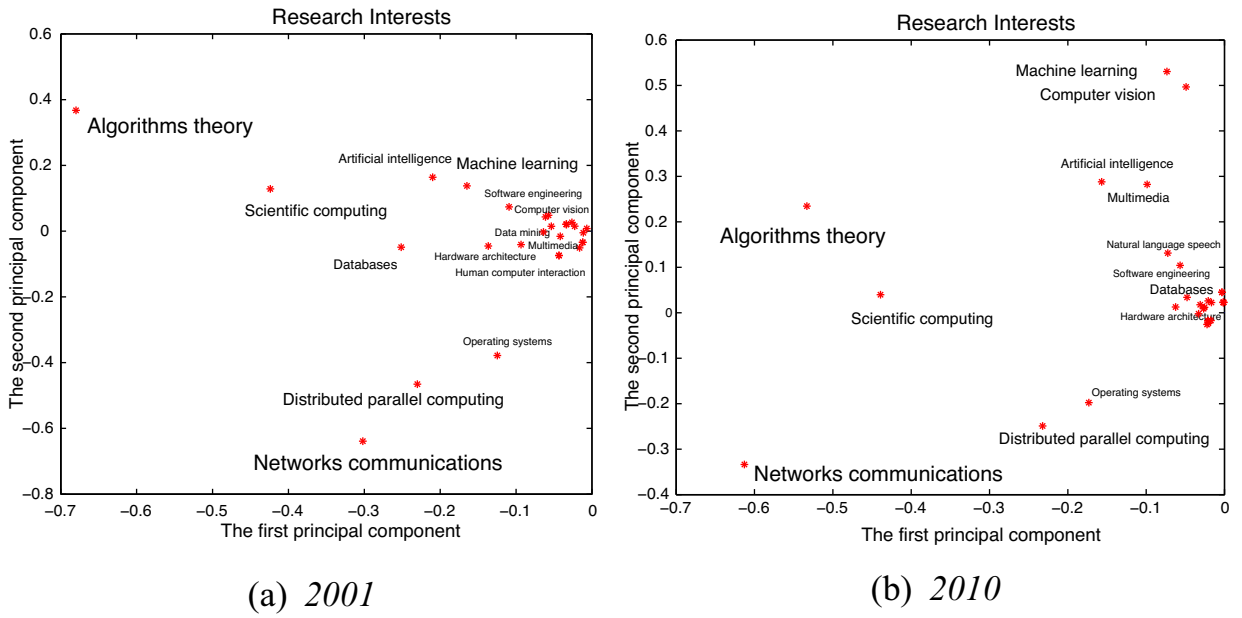


Fig. 4. The statistics of outstanding scientists' *L*-factors. Ten top scientists' *L*-factors in recent 10 years are compiled to a box plot respectively. All the *L*-factors derived from them are larger than 12, which sets a reference of being influential researchers in this field in a way.



(a) *2001* (b) *2010*

Fig. 5. The distribution of mainstream research directions of computer science in 2001 and 2010 computed by principal component analysis method (PCA). The X-axis represents the first principal component and Y-axis represents the second principal component. The asterisks in red color represent the research interests displayed in text. The font size of the research interests is proportional to the number of scientists who are working in this direction. The distance of any two research directions represents the degree of interdisciplinarity between these two directions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Furthermore, we could select excellent scientists in computer science with high L factors in a specific period to analyze the evolution of computer science research.

In this section, we will present an experiment that demonstrates the ability of L-Sequence revealing the developing trend of a research discipline. In brief, 50 scientists were randomly selected with respect to $L_{2001} \geq 10$ and and their research interests were collected as well. With this information, we were able to build a matrix of research interests, RIMatrix, in which the rows represent 50 scientists whereas the columns represent all relevant research interests. Particularly, the element of RIMatrix_{ij} was 1 if the scientist i has research interest j; otherwise RIMatrix_{ij} was 0.

Based on RIMatrix, we performed latent semantic analysis to compute the correlation of all research interests. The distribution of research interests was shown in Fig. 5(a) with font size being proportional to the number of scientists who had the corresponding research interests. The same method was applied to select scientists of $L_{2010} \geq 10$. The popular research interests of 2010 were shown in Fig. 5(b).

From Fig. 5(b), it is clear that 'Algorithms theory', 'Scientific computing' and 'Networks communications' remain as mainstream research directions within recent 10 years. Interestingly, 'Networks communications' has become increasingly popular in 2010 compared to 2001. 'Database' is a rather mature research field and therefore has become less attractive in 2010 than 2001. Moreover, 'Computer Vision' and 'Multimedia' have been gradually attractive for researchers from 2001 to 2010. Noteworthy, some research interests are closely related such as 'Computer Vision' and 'Machine Learning'.

4. Summary

L-Sequence uses a series of measuring factors along with a scientist's career span to reflect his/her research development process and performance. A series of process variables over time can provide an accurate and sufficient description of a scientist's dynamic research trajectory, consistent with the results of peer review and in good accordance with human cognition. To this end, L_α, the sum of all L-Sequence factors, could be used to indicate a researcher's life-long academic performance as h-index or g-index does. However, it could achieve better discriminative power than h- and g-index. Moreover, L-Sequence could not only be reshaped under different evaluating context but also discover research trends at different time period.

Furthermore, another merit of the present work is that yearly citation performance of h-index sequence was thoroughly studied on a large scale depending on our designed computer program which could automatically extract citation counts, citing papers and publication time of the citing papers. This is significant because prior works were severely limited by the dataset collection. In our future work, we hope to establish a large dataset where the sequence information can be easily accessed so as to facilitate informetricians to study various innovative sequences as academic indicators.

References

Abramo, G., & DAngelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics*, *87*, 499–514.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). *H*-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*, 273–289.

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, *436*, 900.

Egghe, L. (2006a]). An improvement of the *h*-index: The *g*-index. *ISSI Newsletter*, *2*, 8–9.

Egghe, L. (2006b]). Theory and practise of the *g*-index. *Scientometrics*, *69*, 131–152.

Egghe, L. (2009a]). Comparative study of *h*-index sequences. *Scientometrics*, *81*, 311–320.

Egghe, L. (2009b]). Mathematical study of *h*-index sequences. *Information Processing and Management*, *45*, 288–297.

Egghe, L. (2010). The hirsch index and related impact measures. *Annual Review of Information Science and Technology*, *44*, 65–114.

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*, 16569.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, *13*, 261–276.

Liang, L. (2006). *H*-index sequence and *h*-index matrix: Constructions and applications. *Scientometrics*, *69*, 153–159.

Liu, Y., & Rousseau, R. (2008). Definitions of time series in citation analysis with special attention to the *h*-index. *Journal of Informetrics*, *2*, 202–210.

Wu, J., Lozano, S., & Helbing, D. (2011). Empirical study of the growth dynamics in real career *h*-index sequences. *Journal of Informetrics*, *5*, 489–497.

Ye, F. Y., & Rousseau, R. (2008). The power law model and total career *h*-index sequences. *Journal of Informetrics*, *2*, 288–297.

Zhang, L., & Glänzel, W. (2012). Where demographics meets scientometrics: Towards a dynamic career analysis. *Scientometrics*, *91*, 617–630.

Zhang, L., Thijs, B., & Glänzel, W. (2011). The diffusion of *h*-related literature. *Journal of Informetrics*, *5*, 583–593.