Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 177 (2007) 3539–3556
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

A computer-aided bibliometric system to generate core article ranked lists in interdisciplinary subjects

Gen Ming Guo

Southern Taiwan University of Technology, Department of Information Management, Lujhu, P.O. Box 102, Taiwan, Kaohsiung 821, Taiwan

Received 27 April 2006; received in revised form 26 February 2007; accepted 28 February 2007

Abstract

Due to the tremendous increase and variations in serial publications, the impact of every peer-reviewed paper on different subjects is varying continually. Domain experts or researchers want to keep track of those latest and highly cited peer-reviewed papers; however they are finding it difficult to update or collect their subject core paper lists regularly and accurately. The evaluation of serial papers for generating and ranking core paper lists on different subjects becomes a very challenging task for scholars or librarians. Therefore, a computer-aided bibliometric system (CABS) was developed to generate a core article ranked list automatically. Four indicators – subject reference cited counts, subject total cited counts, subject reference period impact and subject reference cited history – were proposed to generate a subject core article ranking list. Seven different subjects including E-commerce, Data Mining, Supply Chain, Image Processing, Enterprise Resource Planning, Microarray and Expert Systems were used as samples. The turning point (TP) was proposed to determine the core article area in the paper citation analysis. The TP patterns observed were that all TPs had the same rate for four different subjects. The usage of TP patterns can also be used to verify the experimental results. This study provides experimental evidence to disprove three myths. Myth 1: the top papers on a subject (for instance, the top 10 papers) were all submitted to (S)SCI journals. Myth 2: the highly cited papers (cited counts >4) on interdisciplinary subjects were almost submitted to (S)SCI journals. Myth 3: the articles published in the top journals on a subject would be highly cited.

Keywords: Bibliometrics; Scientometrics; Citation analysis; Paper ranking

1. Introduction and definition of terms

1.1. Introduction

The tremendous increase in the number of serial publications and the impact of a journal paper on multidisciplinary subjects varies continually. As a result, faculties or scholars are finding it difficult to collect core paper lists on their research subjects regularly and accurately. Librarians are finding it increasingly difficult to maintain a current serial collection, which is suitable for interdisciplinary subjects. The evaluation of

E-mail address: sambuela@gmail.com

0020-0255/$ - see front matter
doi:10.1016/j.ins.2007.02.043

peer-reviewed papers for generating and ranking core paper lists on multidisciplinary subjects becomes a very challenging task for university faculties, scholars or librarians.

As a result, a computer-aided bibliometric system (CABS) and several ranking indicators were proposed in order to generate a ranked list of core articles on interdisciplinary subjects for the academic community. Ranking core articles on different subjects to provide a ranked paper list does not only provides high-impact articles to scholars but also reduces the time taken to study all related papers in their subject area. They can have up-to-date knowledge about their subjects of interest by reading only milestone papers and important original papers rather than all related articles. An important paper implies that the paper is highly cited. The first original paper implies the first pioneering paper. An important milestone paper implies a published paper, the subject of which is a breakthrough discovery or innovation. Unveiling the hidden patterns in the literature citations is another objective of this study. After I constructed software tools to generate the ranked lists of core articles on different subjects, I attempted to conduct further citation analysis. I discovered some turning point (TP) patterns in interdisciplinary subjects. These patterns could help researchers to evaluate experimental results and determine core peer-reviewed paper lists on a particular subject area. Researchers can benefit from these generated core paper ranked lists as they can collect, pay and download the core papers on their subjects of interest. Scholars may find that this saves more time as they can utilize the time they would have spent randomly looking for many related papers and then studying them to simply study classic, milestone, or the latest hot papers.

In terms of related works, NEC laboratories America, Inc. has the "CiteSeer", search engine that searches scholarly literature [1]. Google launched "Google Scholar" on November 1, 2004, which searches for peer-reviewed papers, books and technique reports. As far as CiteSeer is concerned, its subject article ranking method is based on the citation counts that are calculated from its own database. The disadvantage of CiteSeer is that a pioneering paper or milestone paper would be ranked behind the highly cited articles, and there are fewer opportunities for the latest hot papers to be ranked higher. "Google Scholar" is based on the Page Rank algorithm employed by "Google Web" [3,13]. It not only calculates the cited counts but also analyzes the citation network in order to find the root web pages or first original paper. In this study, I found that the time complexity could be reduced to $O(n\log n)$ if the time attribute was used. Further the experimental result from the indicators proposed in this study was very close to that of "Google Scholar" (correlation factor >0.7). This time complexity that is termed "Big O" is a part of the theory of computation that deals with the resources required during computation to solve a given problem. In addition, the subject period impact indicator (SRPI) will allow not only the pioneer/milestone papers to be ranked ahead but also the latest hot papers. This would avoid researchers to miss some import papers published in the past or present. There are some other related works on journal ranking [2,7,19]. This study focuses on the article ranking instead of journal ranking; therefore, further comparisons between the two are not described.

In addition, this study provides evidences to disprove three myths. (1) Myth 1: the top papers on a subject (for instance, the top 10 papers) were all submitted to (S)SCI journals. (2) Myth 2: the highly cited papers (cited counts >4) on interdisciplinary subjects were almost submitted to (S)SCI journals. (3) Myth 3: the articles published in the top journals on a subject would be highly cited.

The citation relationships of core papers on different subjects were complicated; further, they form a large citation network topology. Therefore, four presentation views were designed to simplify and present these core articles in the ranked paper lists generated using the CABS. (1) A two-dimensional (2D) citation map that shows the citation association for highly cited papers or top ranked articles on two-axis maps (x-axis: time and y-axis: rank); its major contribution is the simplification of complicated citation relationships because it includes the article ranking and time attributes. Only highly cited papers and top ranked papers of each year would be selected and shown in this 2D citation map; (2) A research evolution tree that shows the evolution of research results in the XML tree format for the top ranked papers such as pioneering, original or milestone papers. It can help scholars to understand how techniques or theories develop and improve using this information visualization tool; (3) A citation pipeline that shows the citation relationships of highly ranked papers in detail using pipeline graph. The pipeline graph is one of the popular visualization solutions to present complicated underground connected pipelines such as gas or water pipelines [11]. This graph can provide a geographic location view that other methods lack; (4) A history timeline which shows the original or milestone research works in history via the format of timeline. It is an easy and popular method used in most history textbooks [15]. The timeline was used and adopted to present subject core ranked articles with different time spans. This study shows

that the proposed presentation solutions for ranked papers on interdisciplinary subjects have unique features and special advantages after comparing with the HighWire [5] and Ke-Börner-Viswanath [9] citation maps.

1.2. Definitions

The import terms and phrases used in this article are defined in the following:

(1) Core Paper List
A group of highly cited peer-reviewed articles published in the same subject area. Most core papers would comprise pioneering, classical, milestone or latest hot articles that have been highly cited in the past or present.
(2) Reference vs. Citation
A. Reference
Each peer-reviewed paper would cite and list other sources and materials such as journal papers, patents, textbooks and so on to provide evidences or comparisons in order to describe the differences among the citing related works. These sources that appear at the end of the paper are together called reference or bibliography.
B. Citation
Every peer-reviewed paper could be cited or referenced by other publications. A cited relationship exists between cited or citing papers. The citation is the citing article rather than the cited material.
(3) Subject Core Journal Ranked List vs. Subject Core Article Ranked List
A. Subject Core Journal Ranked List
A group of highly cited journals in a particular subject area are arranged in sequence on the basis of their importance or impact factor.
B. Subject Core Article Ranked List
A group of highly cited peer-reviewed articles in one particular subject area that are sequenced on the basis of their importance or impact situation.
(4) Turning Point
The accumulated citation counts would increase significantly due to the core ranked papers. The TP is to determine which paper is in the core ranked paper list or the border ranked paper list.

2. Materials and methods

2.1. Materials

The citation raw data sets were retrieved from the Thomson Inc. website [16] for the subject article ranking method. Thomson Inc. acquired Institute of Scientific Information (ISI) Inc. in 1992. Thomson Inc. includes the Thomson Scientific, formerly known as Thomson ISI. The two major sources used in this study were the Journal Citation Report (JCR) database and The Web of Science (WOS) database. JCR stores the journal impact factor for every qualified journal, whereas WOS gathers information for the articles published in a journal in the JCR list. Similarly, I have used the citation data of not just the articles but also the journals. Seven topics were selected in this research: (1) Microarray; (2) Expert Systems; (3) Data Mining; (4) Supply Chain; (5) Image Process; (6) Enterprise Resource Planning and (7) E-commerce. These seven interdisciplinary areas selected are from three different academic schools – medicine, engineering and business schools. In general, each article had 10–30 citing and cited references. The time span was from 1975 to 2004.

2.2. Methods

2.2.1. CABS

The Visual C++.Net programming tool was used to develop one system, called the Computer-Aided Bibliometric System (CABS). The source code is available at www.openfind.idv.tw/core [4]. Fig. 1 shows the

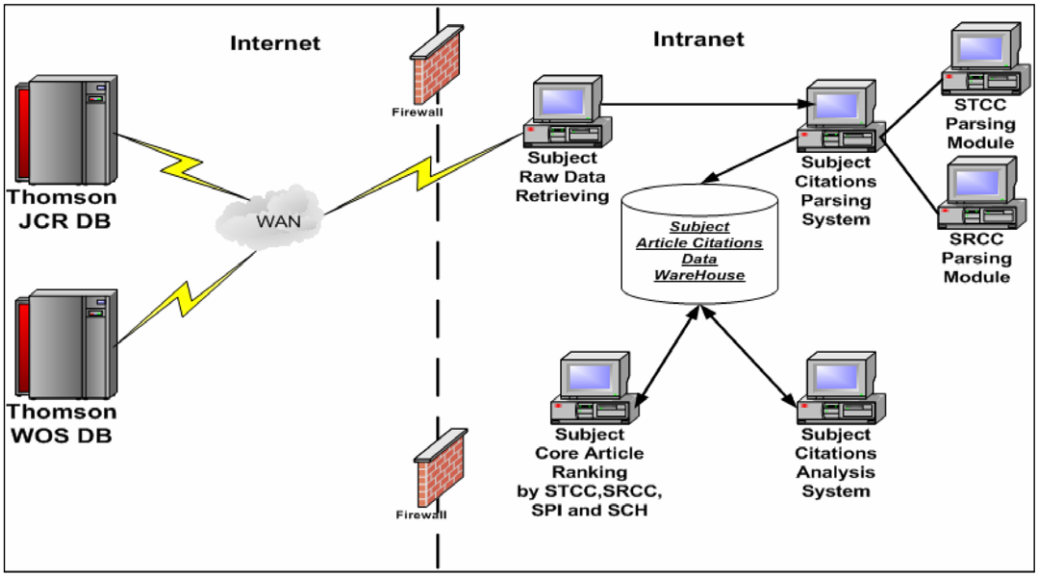

Fig. 1. Network topology of computer-aided bibliometric system (CABS).

network topology architecture of the CABS. Two on-line databases, JCR and WOS, were constructed by Thomson Inc. There are three sub-systems in CABS: the subject citations parsing system, the subject core article ranking system and the subject citation analysis system. Both subject total cited counts (STCC) and subject reference cited counts (SRCC) parsing modules were constructed and embedded in the subject citation parsing system. After the raw citation data of a subject were retrieved, they were parsed by the Subject Total Cited Counts (STCC) and Subject Reference Cited Counts (SRCC) parsing modules and then stored in the data warehouse. Elimination of author self-citation (Section 3.1) was done using the SRCC parsing module. Mapping every journal to Journal Citation Report (JCR) in order to obtain the journal impact factor (JIF) was completed by the subject citation parsing system also. Subsequently, STCC period impact (STPI), STCC cited history (STCH), SRCC period impact (SRPI) and SRCC cited history (SRCH) methods were used to calculate and rank every article from interdisciplinary subjects. The subject citation analysis system assists the TP analysis. The explanations for the STPI/STCH/SRPI/SRCH methods and the turning pointing analysis are described in Sections 2.2.2 and 3.2.

2.2.2. Four ranking indicators

In order to generate the core article ranked list for interdisciplinary subjects, four indicators were proposed. They are SRCC, STCC, SPI (SRPI and STPI) and SCH (SRCH and STCH). There are two major differences between $SRCC_k$ (Subject Reference Cited Counts) and $STCC_k$ (Subject Total Cited Counts). The first difference is the scope of the dataset. $SRCC_k$ (Formula 1) is used to parse the references of retrieved article and then calculate the cited counts from the retrieved dataset. $STCC_k$ (Formula 2) generates the total cited counts from the WOS database. The dataset covers the documents contained in the WOS database. The time period extended from 1975 to 2004. The second difference relates to an author's self-citations, which cannot be removed from the sample set. $STCC_k$ itself does not subtract this self-citation as $SRCC_k$; however, $SRCC_k$ is useful for limited sampling and survey. The results of the query are not tagged in a way that allows a programmatic process to subtract self-citations. Therefore, the limitation of the sources/materials is that the self-citation cannot be removed because of the sample set format. Nevertheless, $SRCC_k$ can provide scholars a bird's eye view of their topics of interest because many articles cite cross-disciplinary papers. In the formula of $SRCC_k$, R_{ij} denotes the cited counts for article k in retrieved particular subject area papers. $SRCC_k$ is cross-disciplinary because R_{ij} includes and calculates all the references for every input paper. These references cite articles on many related subject areas and are not limited to any particular one. $STCC_k$ has an advantage in that it provides scholars a ranked paper list for their subjects of interest. The input article, A_k, is the processing

article among retrieved articles in the CABS. A_i is one of the references of A_k. The cited counts would be subtracted if the target article matches the reference article. $STCC_k$ is a simple indicator without recursive operation. T_i is the total cited counts for the target article in the WOS database. $STCC_k$ is subject focused because T_i only considers and processes the retrieved articles that match the search keyword with the paper title exactly.

$$SRCC_k = \begin{cases} \dfrac{\left(\sum_{j=1}^{n}\sum_{i=1}^{m} R_{ij}\right)}{Nt}, & \text{if } A_k \neq A_i \\ \dfrac{\left(\sum_{j=1}^{n}\sum_{i=1}^{m} R_{ij} - \sum_{j=1}^{n}\sum_{i=1}^{m} S_{ij}\right)}{Nt} & \text{otherwise} \end{cases} \tag{1}$$

where

A_k	the input article
A_i	the input article's reference
i	article' reference ($i = 1,2,\ldots,m$)
j	query results ($j = 1,2,\ldots,n$)
R_{ij}	cited counts for article k in retrieved particular subject area papers
S_{ij}	self-citation counts
Nt	the amounts of query results returned in the type of original article (t = type of article: Ex: original article, review article, news or letter)

$$STCC_k = \sum_{i=1}^{p} T_i/Nt \tag{2}$$

where

T_i	Cited Counts for Article k in Thomson Inc.'s Web of Science Database
p	Total Citing Papers in paper citation database (Ex: Web of Science database)
Nt	The Amounts of Query Results Returned in the Type of Original Article (t = Type of Article: Ex: Original Article, Review Article, News or Letter)

$SRPI_k$ (Formula (3.1)) is an extension of SRCC. The purpose of this factor is to allow new important papers to have a better chance of being ranked highly. The major difference between $STPI_k$ and $SRPI_k$ is that $STPI_k$ is an extension of STCC and is therefore suitable for the dataset from STCC, whereas $SRPI_k$ is designed for SRCC. As for the number of 0.01 in Formula (3.1), the purpose is to avoid the case when $SRCC_k$ or $STCC_k$ is equal to zero.

$$SRPI_k = \frac{(SRCC_k + 0.01)}{Nt}/(YR + 1 - PY) \tag{3.1}$$

$$STPI_k = \frac{(STCC_k + 0.01)}{Nt}/(YR + 1 - PY) \tag{3.2}$$

where

YR	Current Year (ex: 2005)
PY	Published Year (ex: 1987)

$SRCH_k$ (Formula (4.1)) is designed to filter out the classic papers in history, primarily for the senior researchers or authors of review papers. $STCH_k$ (Formula (4.1)) is very similar to $SRCH_k$. This factor is an extension of STCC and is suitable for the STCC dataset.

$$SRCH_k = SRCC_k(YR + 1 - PY) + \frac{1}{(YR + 1 - PY)} \tag{4.1}$$

$$STCH_k = STCC_k(YR + 1 - PY) + \frac{1}{(YR + 1 - PY)} \tag{4.2}$$

3. Results and discussions

3.1. Author self-citations

Author self-citation was a well-known noise in the journal citation analysis [8,12,17]. Through the CABS, the summary data for author self-citation in interdisciplinary subjects are shown in Table 1. The three-step processes followed to generate the results in Table 1 were:

1. Preparation of the raw data sets of the seven subjects to be used as the input to the CABS.
2. Value generation – the CABS matches the author name with the cited references to that author for each article in the data set. The count is used to generate the value for the articles in which the author cites himself or herself.
3. Calculation – the algorithm generates the self-cited articles-to-articles input quantity (for instance, SA ratio = 283 self-cited articles/500 articles) and the self-cited counts-to-total-cited counts ratios (for instance, ST ratio = 1501 self-cited counts/8758 total cited counts) in order to generate the statistical summary data.

The SA ratios for all seven subjects are approximately 0.6. A pattern is observed. The ST ratio (self-cited counts-to-total-cited counts) varies from 0.06 to 0.17 and the average is 0.1. These data again reveal that most authors often cite their own papers. Owing to the high SA ratio (average >0.5), author self-citation would cause a serious bias in the citation analysis. Therefore, the self-citation should be subtracted from the total citations in order to obtain an accurate impact factor for different journals or articles.

3.2. The turning point pattern

In the citation analysis, I found that all subjects exhibit TP pattern. In this article, the Turning Point (TP) implies that the accumulated citation counts would increase by the ranked papers after ranking a paper by either STCC or SRCC. A TP emerges when the cited counts on the y-axis do not grow significantly after the geometry analysis was conducted. The procedure that I adopted to obtain the TP was as follows. First, the ranked articles were divided into 14 segments, each having equal article counts. Fourteen segments is the default setting in Microsoft Office Excel. Second, I observed where the TP was located by using the geometrical method as shown in Fig. 2a. The geometric algorithm/procedure is illustrated below. All the TPs in Fig. 2 and Table 2 were located at the boundary of the 3rd segment in the 14 segments (TP site: 0.21) for four random selected subjects (Microarray, E-commerce, Expert Systems and Data Mining) by the STCC indicator. The TC Ratios (TP count-to-total count) are all approximately 0.2. All TP angles are approximately 70°. (TP angle: the acute angle for the TP) Using the SRCC method, all TPs in Fig. 3 and Table 3 from four subjects (Microarray, E-commerce, Expert Systems and Data Mining) were all located at the boundary of the 1st segment in the 14 segments (TP site: 0.07). The values of both TC ratios and TP angles from SRCC are close to the experimental results from the STCC indicator. The patterns of the TC ratios and TP angles are 0.2° and 70°, respectively. These TPs/TC Ratios/TP angles patterns would be helpful in determining the core article area or evaluating the experimental results in interdisciplinary subjects. All the subjects chosen in this study have the same value for TP; therefore, I expect that TP could be extended to other subjects too. The librarian

Table 1
The author self-citation ratio table in seven subjects

Subject	Self-cited articles	Article quantity	Self-cited counts	Total cited counts	SA ratio	ST ratio
E-Commerce	283	500	745	11943	0.6	0.06
Data mining	282	500	861	7477	0.6	0.12
Supply chain	224	500	592	8864	0.5	0.07
Image process	310	500	1501	8758	0.6	0.17
ERP	88	178	229	4153	0.5	0.06
Microarray	232	500	627	9635	0.5	0.07
Expert systems	355	500	1116	9697	0.7	0.12

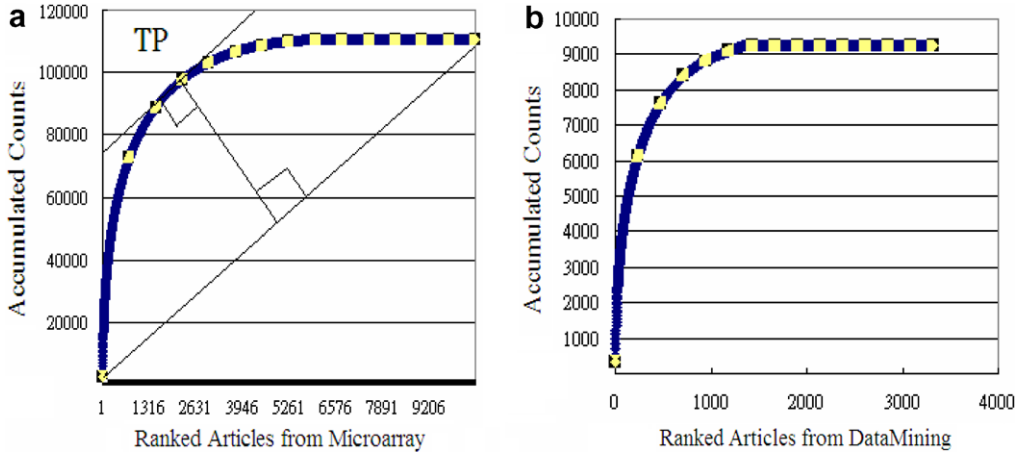

Fig. 2. The citation chart and turning point by STCC (subject total cited counts) method in interdisciplinary subject areas. (a) Microarray and (b) data mining.

Table 2
Turning points of four subjects from STCC (subject total cited counts)

	TP site	TP count	Total count	TC ratio	TP angle
Microarray	0.21	2397	11,348	0.21	75
E-Commerce	0.21	2687	14,002	0.19	70
Expert systems	0.21	1767	10,765	0.16	71
Data mining	0.21	1697	8712	0.19	68

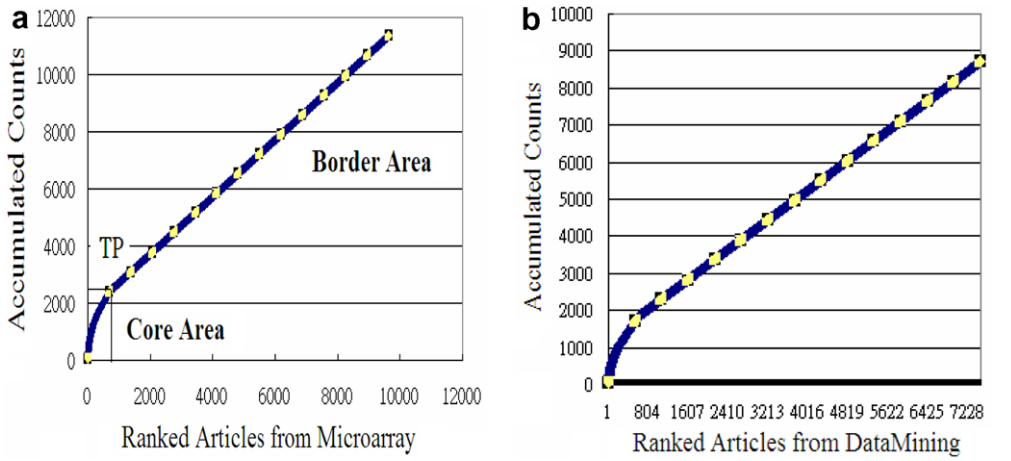

Fig. 3. The citation chart and turning point by SRCC (subject reference cited counts) in different subject area. (a) Microarray and (b) data mining.

may change the current e-journal procurement policy for not subscribing publishers' complete databases and only subscribe to the highly cited papers. This could save more money and increase the access rate for the subscribed electronic papers. Researchers can also refer to this TP pattern to pay and download core papers from commercial databases in order to collect and study the core ranked papers in their subjects of interest.

Table 3
Turning points of four subjects from SRCC (subject reference cited counts)

	TP site	TP count	Total count	TC ratio	TP angle
Microarray	0.07	2397	11348	0.21	75
E-Commerce	0.07	2687	14002	0.19	70
Expert systems	0.07	1767	10765	0.16	71
Data mining	0.07	1697	8712	0.19	68

```
//*The Algorithm of Geometry to Calculate Turning Point.*//
//Please refer to Fig. 2a
p0 = the leftmost point;
p1 = the rightmost point;
for all points p2 from right to left {
        Calculate SLOP (p0, p2);
   repeat {
        Move p1 left;
        Calculate SLOP (Find the tangent lines that pass through p1);
   } until SLOP (Tangent line to pass through p1) = SLOP (p0, p2)
        line0 = draw a line to pass through point p1;
        line1 = draw a line to pass through point (p0, p2);
        Determine if two lines that intersect at a 90 degree angle and line1 is divided into two equal parts.
   if YES {Output the p1}
} //End of for
```

3.3. The myths

Due to the JCR provided by Thomson Inc., many scholars have either been guided or affected by its SCI (Science Citation Index) or SSCI (Social Science Citation Index). However, this study provides evidences to disprove three myths. (1) Myth 1: the top papers on a subject (for instance, the top 10 papers) were all submitted to (S)SCI journals. (2) Myth 2: the highly cited papers (cited counts >4) on most subjects were submitted to (S)SCI journals. (3) Myth 3: the articles published in the top journals on a subject (for instance, top three journals) would be highly cited.

With regard to Myth 1, I differentiate three types of citation models based on the hierarchical clusters in Fig. 4(a and b). Most citation models of seven subjects belong to Type II. In Type II, the (S)SCI citation per-

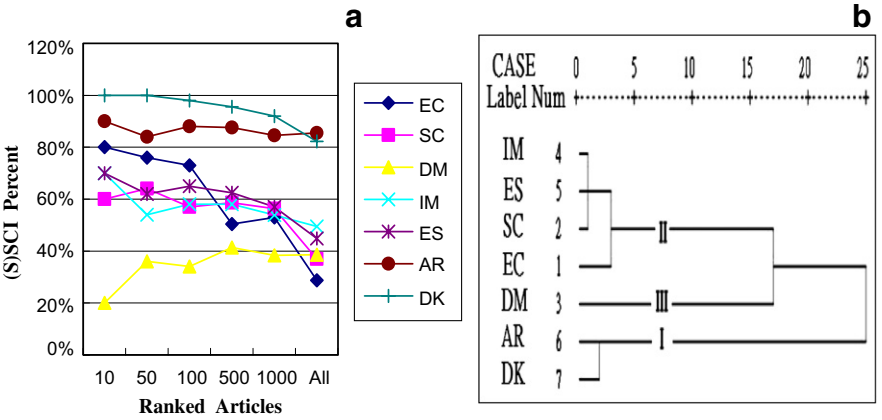

Fig. 4. Three types of citation models were obtained by clustering ranked articles and (S)SCI percent in seven subjects areas. (a-Left): The (S)SCI percent in the ranked article citations. (b-Right): The hierarchical cluster results expanded from Fig. 4a.

centage is very stable. After I ranked papers using the SRCC method, I observed that for myth 1, the citation percentage of only Type I was more than 80%. Type I of both Microarray (AR) and Docking (DK) are in the subject area of biomedicine. In generally, peer-reviewed papers published in the field of biomedicine would have a higher citation frequency than those of other fields. For instance, the top 25 highly cited journals in Journal Citation Report are all almost in the subject area of biomedicine [14,16]. For type III, the citation percentage of Data Mining is less than 20%. Further investigations have to be conducted to determine why the Data Mining (S)SCI Ratios were so less and why so many top papers were only submitted to conferences rather than (S)SCI or non-(S)SCI journals.

With regard to myth 2, I have observed similar types and models in Fig. 5(a and b) and Fig. 4(a and b). In seven subjects, most (S)SCI percentages from the top articles, that is, the articles that were cited at least four times are also less than 80%. The percentage of both type II and III are below 80% as well. These pieces of evidence show that many highly cited articles were not submitted to (S)SCI journals. Although the reasons for this need to be explored further, this situation is very common for ranked or clustered articles on interdisciplinary subjects.

For myth 3, MIS Quarterly (MISQ), Information Systems Research (ISR) and Decision Support Systems (DSS) are three top journals for the E-commerce subject area or the department of information management in universities [6,18,19]. However, E-commerce still has many articles in the three top journals that were not cited by any other articles in a different time span. The exact citation counts for every article are illustrated in Fig. 6 and can be retrieved from www.openfind.idv.tw/core.

For Myth 1 and Myth 2, I formed two hypotheses and tested them using the t-test. I assumed that the data set was obtained from a normal distribution. In Myth 1, the null hypothesis (Formula 5) was that less than 60% of the top 10 papers on different subjects were submitted to (S)SCI journals on an average. The t-value was 1.0255 ($t-\text{value} = (0.7-06)/(0.258/\sqrt{7})$) and the α was set to 0.05. H_0 was not significant because of t-value $< t_{6,0.05}$. Therefore, the null hypothesis cannot be rejected and Myth 1 was disproved. As for Myth 2, the H_0 (Formula 6) was that less than 50% of the highly cited papers (cited counts >4) on different subjects were submitted to (S)SCI journals on an average. The t-value was 1.838 ($t-\text{value} = (0.648-0.5)/(0.213/\sqrt{7})$) and the α was set to 0.05. The null hypothesis was not significant because of t-value $< t_{6,0.05}$. Therefore, H_0 cannot be rejected and Myth 2 was disproved.

$$\begin{cases} H_0 : \mu \leqslant 0.6 \\ H_1 : \mu > 0.6 \end{cases} \tag{5}$$

$$\begin{cases} H_0 : \mu \leqslant 0.5 \\ H_1 : \mu > 0.5 \end{cases} \tag{6}$$

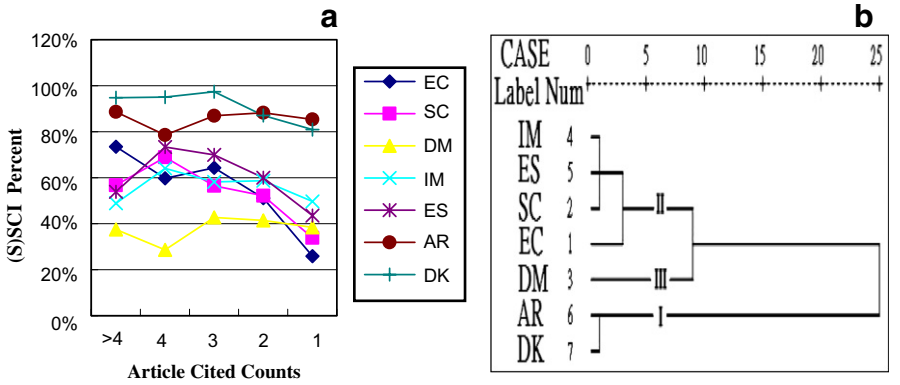

Fig. 5. Three types of citation models were obtained by clustering cited counts and article cited counts in seven different subjects areas. (a-Left) The (S)SCI percent in citation frequency. (b-Right) The hierarchical cluster results extended from a.

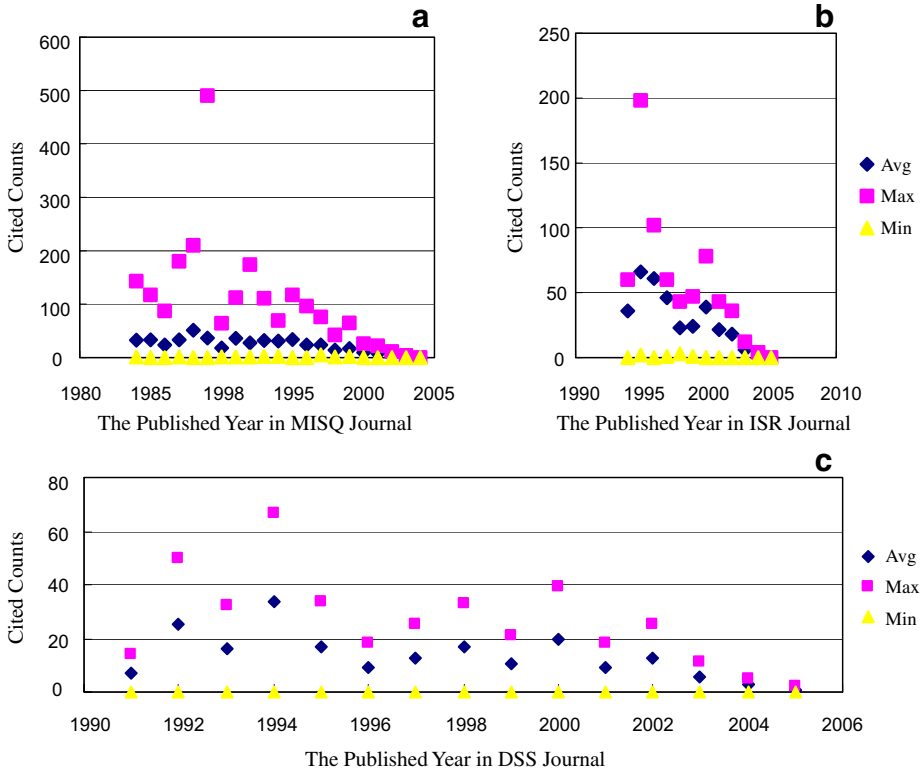

Fig. 6. Three top journals in the E-commerce subject area: (a: upper-left figure) The citation chart of MISQ Journal. (b: upper-right figure) The citation chart of ISR Journal. (c: lower figure) The citation chart of DSS Journal.

3.4. Comparison of different indicators

3.4.1. Citation network analysis

One citation network is illustrated in Fig. 7 in order to explain and compare the proposed indicators. In this figure, microarray is used as an example. Time is represented on the x-axis. Each alphabetic word represents one paper. The arrow indicates a citation link. In this citation network, the problem faced by both STCC and SRCC would be to determine which one should be ranked No. 1 in Fig. 7. This is because both 'D' and 'A' papers would obtain equal weights in this case. In this situation, the first original paper would not be ranked No. 1. This is the reason why both STCC and SRCC are unsuitable for generating a ranking list for historians or author(s) of review papers. However, they could be extended to other purposes. With regard to STPI or SRPI, the 'D' paper will have a higher score than the "A" paper to be ranked No. 1, even though both the "D" and "A" papers have six link-out papers. Thus, new good papers would have more chances to be listed ahead. This would be very suitable to provide students with one suggestion list of classical and latest hot papers to study while they just enter a new field. STCH and SRCH had been designed to not only filter out highly cited papers but also to search for the first original papers. Therefore, the first original paper "A" would be ranked ahead of a milestone paper "B". In fact, Fig. 7 depicts a real case. After I checked all the papers in the WOS database of Thomson Inc., I found that the paper "A" was the first original paper of Microarray rather than the paper "L". In general, every milestone paper would also cite other papers. But the first original paper would be more highly cited in the early age than the others. Its published time is earlier than other milestone papers. As a result, the attributes of time and cited counts are important attributes in STCH and SRCH indicators. Both the Google and Google Scholar search engines [3,13] use similar ideas as the STCH/SRCH indicator to rank web pages or academic papers. However, a major difference between STCH/SRCH and Google is the time attribute, which is important to paper citations but not for web page citations. The time attribute can reduce the Big O of time complexity to$O(n\log n)$. The page rank formula from

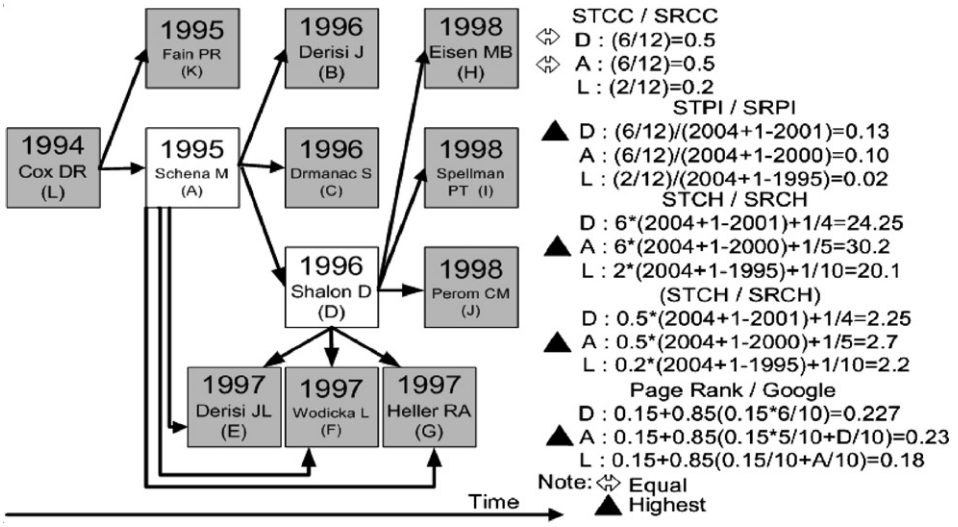

Fig. 7. The journal paper citation network analysis.

Google is "$PR(A) = (1 - d) + d(PR(T1)/C(T1) + \cdots + PR(Tn)/C(Tn))$." To quote from Google paper, page "A" has pages T1···Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. They usually set d to 0.85. Also $C(A)$ is defined as the number of links going out of page A. With regard to Microarray example in Fig. 7, both STCH/SRCH and PR/Google can find the first original paper "A". I have listed STCH/SRCH twice in Fig. 7. This is because the second STCH/SRCH can just use the citation counts to replace the output values from STCC/SRCC and it would save more time and obtain the similar results as the first STCH/SRCC.

3.4.2. Strength and weakness

Each indicator has been designed for a different purpose and audience (Table 4). Both SRPI/STPI and SRCH/STCH were extended from the results of SRCC and STCC. Due to the limitation of the raw data set, STCC can not eliminate author self-citations. This is the weakness of STCC. However, STCC can produce a customized paper ranking list for each journal. For example, STCC can rank all papers that had been published in one journal in every time span; this could not be achieved by SRCC. The advantage of SRCC was that it could generate one paper ranking list for several related journals at one time, which would be helpful to provide a bird's eye view of the research topic. In Fig. 7, it is shown that a new published paper has chances to be ranked higher by the SRPI/STPI methods. Therefore, it would be suitable to provide a ranking list with the latest hot and classical papers. In Fig. 7, SRCH/STCH can filter out the original and milestone papers. This would help historians or authors of review papers to observe the research developments in history or write the review papers.

Table 4
The strengths, weaknesses, audience and purpose analysis for different indicators

	SRCC	STCC	SRPI	STPI	SRCH	STCH
Strength	No self-citation A bird's view	Focus Impact factor	Hot topic Normalize time	Hot topic Normalize time	Original Mile stone	Original Mile stone
Weakness	Focus Time complexity	Self-citation A bird's eye View	Fever-phenomenon	Fever-phenomenon	Blooming-research	Blooming-research
Audience	Surveyor	Senior scholar	Beginner, general Student	Beginner, general Student	Senior Historian	Senior Historian
Purpose	Understand the overall situation	Explore in-depth	Filter out the hot research works	Filter out the hot research works	Trace evolution	Trace evolution

Table 5 shows a comparison of related indicators. The advantage of SRCH was that an author's self-citation was subtracted in this method. Self-citations could cause serious bias and noise in the citation analysis. This has been discussed in Section 3.1. Both SRCH and STCH can filter out the classical papers in the same manner as 'Google Scholar'. However, SRCH/STCH can reduce the time complexity to O($n\log n$) because the time variable was used. Articles such as original papers or review papers were considered to be the inputting sources. This can avoid inputting sources from news, letters or comments articles which could be the noise factors. With regard to the CiteSeer, its ranking method was based only on the factor of citation counts. In this way, the first original paper can not be filtered out and ranked ahead if milestone papers have higher citation counts.

3.4.3. Correlation and distance

The correlation coefficient [10] was used to calculate the distances between the different ranking methods. It is a popular formula and factor in the statistical methods to calculate the correlation between two series of numbers. Microarray was the analysis example used in this study. All the correlation factors between methods are shown in Table 6. They are highlighted in four shades and clustered into four groups. SRCC/SRPI/SRCH had a higher correlation factor than STCC/STPI/STCH. SRCC/SRPI/SRCH had a shorter correlation distance to "Google Scholar" than STCC/STPI/STCH. The correlation factors in the left hand upper corner shown in a light gray surpassed the other areas in Table 6. The right hand lower corner had the lowest factors. The mid-area, shown in black, also had low factors. To summarize, the rule was "SRCC/SRPI/SRCH > Scholar/SRCC/SRPI/SRCH > STCC/STPI/STCH > Overlapping of SRCC/SRPI/SRCH and STCC/STPI/STCH > Scholar/STCC/STPI/STCH". The main reason for the higher correlation factor of Scholar/SRCC/SRPI/SRCH as compared to that of Scholar/STCC/STPI/STCH was that STCC/STPI/STCH only calculates data sets from the articles published in the SCI/SSCI journals list. Both the "Google Scholar" search engine and SRCC/SRPI/SRCH did not limit their data sets to SCI/SSCI journals. In particular, some highly cited articles were not submitted to SCI or SSCI journals in interdisciplinary subjects. This has been explained in Section 3.3 (Myths). As for the overlapping area of SRCC/SRPI/SRCH and STCC/STPI/STCH, lower correlation factors were expected because the sizes of the data set from SRCC and STCC were not exactly equal. Using their own extended methods, both SRCC and STCC are able to obtain more than 0.7 and 0.9 correlation factors, respectively, SRCC/SRPI/SRCH has more than 0.72 correlation factors with "Google Scholar". Therefore, the experimental results show that SRCC, SRPI and SRCH were measured to be at an acceptable level. In Table 7, the top 50 overlapping papers were listed for the Microarray subject area.

Table 5
A comparison of different subject article ranking indicator

	SRCH	STCH	Google scholar	CiteSeer
Purpose	Peer-reviewed Paper ranking	Peer-reviewed Paper ranking	Paper and textbook ranking	Peer-reviewed paper ranking
Strength	No self-citation	Speed, article type and trace paper	Trace classical paper	Time complexity
Weakness	No search function	No search function	Source type (Ex: Letters)	Trace the evolution
Time complexity	O($n\log n$)	O($n\log n$)	O(n^2)	O($n\log n$)

Table 6
Distances between indicators in the field of microarray

	SRCC	SRPI	SRCH	STCC	STPI	STCH	Scholar
SRCC	1	*0.99*	*0.98*	**0.428**	**0.327**	**0.274**	0.731
SRPI	0.99	1	*0.96*	**0.43**	**0.357**	**0.25**	0.73
SRCH	0.98	0.96	1	**0.436**	**0.294**	**0.32**	0.728
STCC	0.428	0.43	0.436	1	*0.793*	*0.78*	*–0.31*
STPI	0.327	0.357	0.294	0.793	1	*0.3*	*–0.42*
STCH	0.274	0.25	0.32	0.78	0.3	1	–0.21
Scholar	0.731	0.73	0.728	–0.31	–0.42	–0.21	1

Table 7
Top 50 overlapping papers in the subject area of microarray

Article title	STCC	STPI	STCH	SRCC	SRPI	SRCH
Cluster analysis and display of genome-wide expression patterns	1	1	2	1	1	2
Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray	2	4	1	2	5	1
Exploring the metabolic and genetic control of gene expression on a genomic scale	3	5	3	5	6	4
Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling	4	2	5	4	2	8
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring	5	3	4	3	3	5
Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization	6	9	8	18	31	23
Molecular portraits of human breast tumours	7	7	12	12	8	19
Use of a cDNA microarray to analyse gene expression patterns in human cancer	8	14	6	11	17	6
The transcriptional program in the response of human fibroblasts to serum	9	10	11	17	18	27
Parallel human genome analysis: microarray-based expression monitoring of 1000 genes	10	22	7	15	35	11
Significance analysis of microarrays applied to the ionizing radiation response	11	8	19	7	4	22
Tissue microarrays for high-throughput molecular profiling of tumor specimens	12	15	9	10	12	10
Gene expression profiling predicts clinical outcome of breast cancer	13	6	29	27	13	95
The transcriptional program of sporulation in budding yeast	14	19	13	51	60	49
Printing proteins as microarrays for high-throughput function determination	15	11	17	13	10	21
Genome-wide expression monitoring in *Saccharomyces cerevisiae*	16	25	10	30	25	51
Systematic variation in gene expression patterns in human cancer cell lines	17	12	20	22	19	43
Molecular classification of cutaneous malignant melanoma by gene expression profiling	18	13	21	23	20	45
Genomic expression programs in the response of yeast cells to environmental changes	19	21	24	83	82	126
A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization	20	48	16	20	44	16
Distinctive gene expression patterns in human mammary epithelial cells and breast cancers	21	30	23	26	32	34
Global analysis of protein activities using proteome chips	22	16	36	56	47	75
Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications	23	17	37	46	30	89
Integrated genomic and proteomic analyses of a systematically perturbed metabolic network	24	18	38	155	117	293
A gene expression database for the molecular pharmacology of cancer	25	26	30	64	50	83
Gene-expression profiles in hereditary breast cancer	26	20	40	44	29	87
Discovery and analysis of inflammatory disease-related genes using cDNA microarrays	27	49	18	31	48	30
Comparative genomics of BCG vaccines by whole-genome DNA microarray	28	34	25	160	203	166
Gene expression profiles of laser-captured adjacent neuronal subtypes	29	36	26	66	63	65
High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays	30	45	22	53	71	52
Knowledge-based analysis of microarray gene expression data by using support vector machines	31	31	34	88	56	160
Delineation of prognostic biomarkers in prostate cancer	32	27	48	47	73	36
Genome-wide analysis of DNA copy-number changes using cDNA microarrays	33	43	31	34	38	40
Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations	34	35	39	28	24	50
A gene-expression signature as a predictor of survival in breast cancer	35	23	84	138	70	377
Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion	36	39	43	140	139	183

(*continued on next page*)

Table 7 (*continued*)

Article title	STCC	STPI	STCH	SRCC	SRPI	SRCH
Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation	37	24	87	40	15	122
Drug target validation and identification of secondary drug target effects using DNA microarrays	38	66	28	57	39	112
Genome-wide location and function of DNA binding proteins	39	42	46	201	209	287
Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks	40	33	64	62	96	48
Microarrays: biotechnology discovery platform for functional genomics	41	72	32	55	133	31
Coordinated plant defense responses in Arabidopsis revealed by microarray analysis	42	46	47	482	519	659
Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation	43	47	49	3260	3290	3953
Vascular channel formation by human melanoma cells in vivo and in vitro: Vasculogenic mimicry	44	62	41	2370	3163	2658
Prediction of central nervous system embryonal tumour outcome based on gene expression	45	28	102	85	83	128
Yeast microarrays for genome wide parallel genetic and gene expression analysis	46	93	27	78	127	61
Complete genome sequence of *Salmonella enterica* serovar typhimurium LT2	47	38	70	2164	2189	3778
Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats	48	65	42	531	676	568
The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma	49	29	106	186	145	349
Orchestrated transcription of key pathways in Arabidopsis by the circadian clock	50	50	52	444	495	609

3.5. Graphical data presentation

Due to the complicated paper citation relationships among the core ranked articles on a subject generated by the CABS, four visualization solutions were proposed to simplify the citation relationships topology of the subject's ranked papers and to visualize the filtered articles of subjects on the basis of several factors such as time, quantity and so on. They are as follows: (1) A 2D citation map; (2) research evolution (XML) tree; (3) citation pipeline and (4) history timeline. I have tried to use a table to list and compare the citation map of HighWire's [5] and Ke-Börner–Viswanath's [9]; this comparison is provided in Table 8. The staff of the University of Stanford proposed an expandable citation map to present a citation network on the HighWire online journal database portal. They represent the first original article in blue and the highly cited articles in yellow. This map can be expanded to up to 150 nodes. This map has a disadvantage in that the arrows could be mixed together if the citation map is large and complicated. Therefore, this solution is not suitable to present a large citation network. The advantage of Ke-Börner–Viswanath's citation visualization is that color and size are used to present the cited counts of each node/article. Its disadvantage is that it did not consider a time variable. Further, it is necessary to design and arrange the node's location carefully because the nodes with a large radius would overlap the other small nodes on the citation map.

In this study, the proposed 2D citation map can take full advantages of the time and ranked number attributes. In Fig. 8, the x-axis represents the number of articles ranked by the STCH indicator and the y-axis represents the published year. The top three articles were selected to be listed for each year. The degrees of the citation input/output were added to each node. It also can provide a research evolution view. However, the expanding view is a disadvantage as it is impossible to expand the nodes.

The research evolution tree (Fig. 9) inherits some good features from the tree data structure, such as an expanded view, node hierarchy, and so on. Due to the hierarchy feature of the tree, it can not only provide a research evolution view but also a topic cluster view. The research evolution tree can also easily expand the nodes based on the timeline. In addition, it is easy to be integrated with XML (Extensible Markup Language) because both have the same tree data structure. After integrating with XML, it could exchange data

Table 8
The strength and weakness of six presentation views

	Ke-Börner–Viswanath	HighWire	Two dimensions	Evolution tree	Pipeline	Timeline
Strength	(1) A bird's eye view (2) Use color and size to present the node 's cited counts	(1) Expand view (2) Use color and size to present article type	(1) Evolution view (2) Attribute of *X*/*Y* axis (3) Input/output degrees	(1) Evolution view (2) Expand easily (3) Integr ate XML (4) Timeline node (5) Cluster view	(1) Timeline node	(1) History view (2) Textbook's reference
Weakness	(1) Time (2) Location	(1) One dimension (2) Huge network	No expanding view	(1) Duplicate node	(1) Complex cross node	(1) No article link

with other research evolution trees easily. However, this would produce duplicated nodes in the tree. This is because some nodes must be added to a different branch. If this is not done, the research branch or category would be incomplete. This is the tradeoff.

A pipeline was used to present the citation network in Fig. 10. However, it is not easy to develop software to generate this citation pipeline because the cross relationships among different nodes is very complicated and can easily to cause confusion. Therefore, the pipeline is not a good visualization tool for citation relationship.

The timeline is a very popular presentation tool in general history books or museums [15]. The most important events were recorded in the timeline graph. In Fig. 11, a timeline graph was applied to the subject area of Microarray to provide a view of the research history timeline. The top three papers were selected and listed in this timeline graph in different time spans. The disadvantage of a timeline graph is that it does not allow

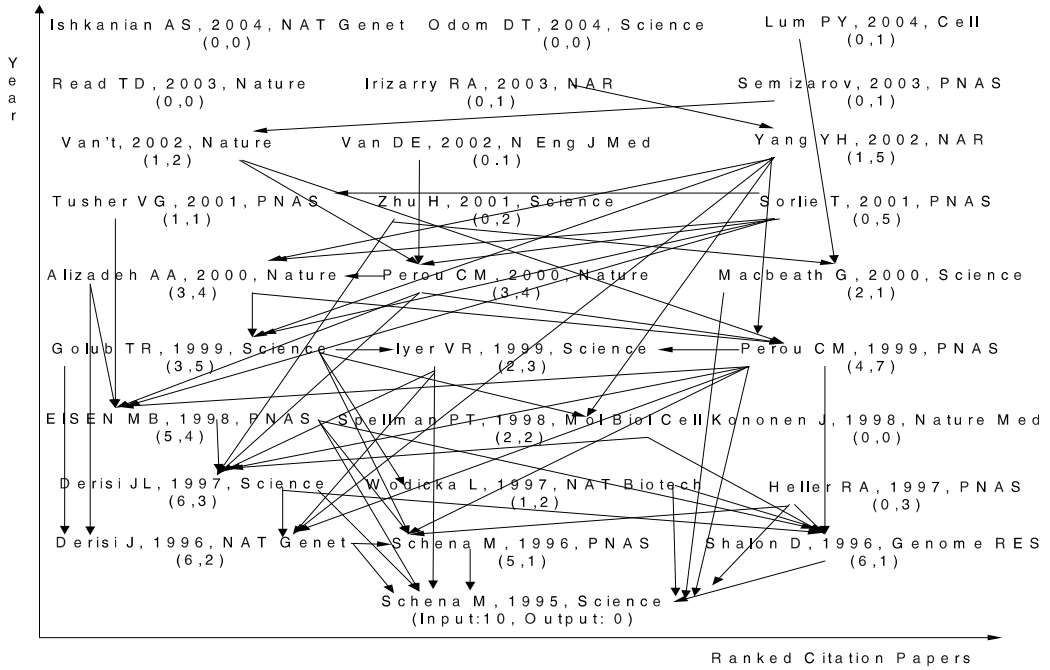

Fig. 8. The two dimensions citation map of microarray.

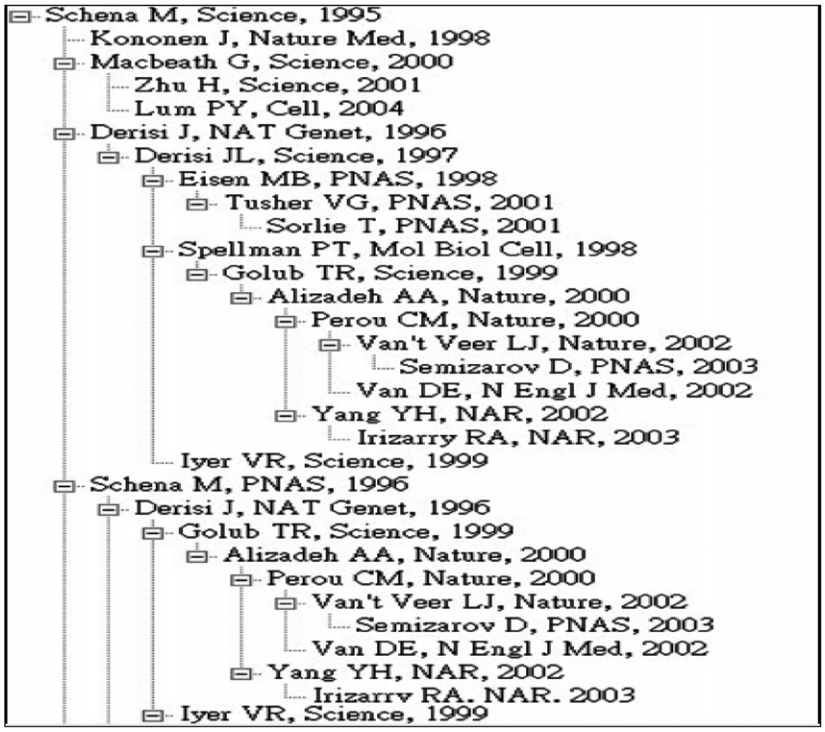

Fig. 9. The research evolution tree (XML tree) of microarray.

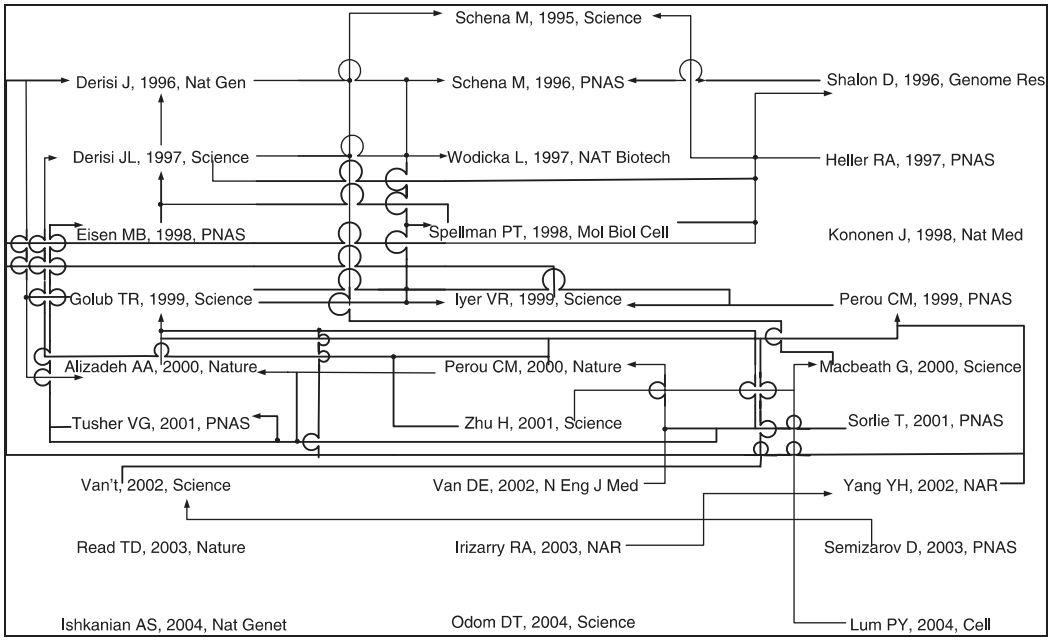

Fig. 10. The citation pipeline in the subject area of microarray.

readers to know the relationships between papers although it can provide a research evolution view. The example used in Figs. 8–11 were all from the same data set and in the subject area of Microarray (Biochip).

Year	Event	Even or Title	First Author	Count
1985	Term first appears	Microarray Electrochemical flow detectors at high applied potentials and liquid-chromatography with electrochemical detection of carbamate pesticides in river water	ANDERSON JL	60
1995	Original Article	Quantitative monitoring of gene expression patterns with a complementary-DNA microarray	SCHENA M	2248
1996	Analyze Cancer	Use of a cDNA microarray to analyze gene expression patterns in human cancer.	DERISI J	862
1997	Metabolic	Exploring the metabolic and genetic control of gene expression on a genomic scale.	DERISI JL	1641
1998	Cluster Analysis	Cluster analysis and display of genome-wide expression patterns	EISEN MB	2493
1998	Cell Cycle	Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization	SPELLMAN PT	890
1998	Tissue array	Tissue microarrays for high-throughput molecular profiling of tumor specimens	KONONEN J	675
1999	Cancer Prediction	Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring	GOLUB TR	1518
2000	Distinct lymphoma	Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling	ALIZADEH AA	1563
2000	Portrait Breast Tumor	Molecular portraits of human breast tumors	PEROU CM	865
2001	Ionizing Radiation	Significance analysis of microarrays applied to the ionizing radiation response	TUSHER VG	696
2001	Proteome Chips	Global analysis of protein activities using proteome chips	ZHU H	393
2002	Predict Breast Cancer	Gene expression profiling predicts clinical outcome of breast cancer	VANT VEER LJ	647
2002	Predict Cancer Survival	A gene-expression signature as a predictor of survival in breast cancer.	VAN DE	288
2003	Bacteria	The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria	READ TD	88
2003	RNAi	Specificity of short interfering RNA determined through gene expression signatures	SEMIZAROV D	79
2004	Complete Human Genome	A tiling resolution DNA microarray with complete coverage of the human genome	ISHKANIAN AS	22

Fig. 11. The history timeline of microarray.

4. Conclusion

A computer-aided bibliometric system (CABS) was designed and constructed in order to generate a subject core article ranking list and assist with further citation analysis. Researchers can pick, download, pay and study the core articles generated from CABS. Scholars may find that this saves more time as they can select and study classic, milestone or the latest hot papers via CABS. It could help avoid searching for many papers randomly and reduce the effort required to survey related works. Four indicators (SRCC, STCC, SPI and SCH) were proposed to generate the core article ranked list in interdisciplinary subjects. Both "Google Scholar" and CiteSeer were used as a benchmark to show that the proposed indicator is at the acceptance level. Four subjects (Microarray, E-Commerce, Expert System and Data Mining) were used as the samples to explore TP pattern. The TP to determine the core articles zone in the subject article ranking list exists in all four subjects. The TP may also exist in multidisciplinary subject areas and could be further extended to other subjects. The research productivity and performance of scholars were often credited by publishing articles on top journals such as (S)SCI journal lists with high impact factor provided by Thomson Inc. This have more honor than to produce the highly cited articles published in the journal without impact factor or high impact factor. This study provides evidences to disprove three myths. Myth 1: the top papers on a subject (for instance, the top 10 papers) were all submitted to (S)SCI journals. Myth 2: the highly cited papers (cited counts >4) on interdisciplinary subjects were almost submitted to (S)SCI journals. Myth 3: the articles published in the top journals on a subject would be highly cited. In addition, due to the complex paper citation network in the real world, four presentation views were designed and used to present the article citation relationships; they are as follows: a 2D citation map; research evolution tree; citation pipeline and history timeline. After comparing with related works such as the HighWire and Ke-Börner–Viswanath citation maps, this study

shows that the proposed presentation solutions have certain unique features and are useful to present the citation data and relationships.

References

[1] Citeseer, 1/29-last access, NEC laboratories America Inc., <www.citeseer.ist.psu.edu/cs>, 2007.
[2] D. Jonathan, T. Narongsak, Ranking the technology innovation management journals, Journal of Product Innovation Management 21 (2) (2004) 123–139.
[3] Google Scholar, Google Inc., <scholar.google.com> (accessed on 23.1.2007).
[4] G.M. Guo, Author supplementary materials. <www.openfind.idv.tw/core> (accessed on 27.1.2007).
[5] Highwire Journal Database, <http://www.highwire.stanford.edu> (accessed on 25.1.2007).
[6] H. Huang, S. Hsu, An evaluation of publication productivity in information systems: 1999 to 2003, Communications of the Association for Information Systems 15 (2005) 555–564.
[7] H. Gaines, A. Minesh, Methods of ranking economics journals, Journal Atlantic Economic Journal 32 (2) (2004) 140–149.
[8] H. Snyder, S. Bonzi, Patterns of self-citation across disciplines, Journal of Information Science 24 (6) (1998) 431–435.
[9] K. Weimao, K. Börner, L. Viswanath, Citation map visualization, <http://www.iv.slis.indiana.edu/ref/iv04contest> (accessed on 19.1.2007).
[10] L. Tian, J.C. Cappelleri, A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: the generalized variable method, Statistics in Medicine 23 (13) (2004) 2125–2135.
[11] M. Balat, The case of Baku–Tbilisi–Ceyhan oil pipeline system: a review, Energy Sources 1 (2) (2006) 117–126.
[12] M. Tsay, Journal self-citation study for semiconductor literature: synchronous and diachronous approach, Information Processing & Management 42 (6) (2006) 1567–1577.
[13] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, WWW7/Computer Network 30 (1–7) (1998) 107–117.
[14] Scibytes, <www.in-cites.com/research/2002/february_18_2002-1.html> 2002.
[15] S.I. Gass, A.A. Assad, An Annotated Timeline of Operations Research: An Informal History, Springer Press, New York, 2004.
[16] Thomson Corporation, ISI web of science and journal citation report. <www.isinet.com> (accessed on 23.1.2007).
[17] Y.C. Shiue, G.M. Guo, A method to build core article ranked lists in interdisciplinary departments and journals, Journal of Educational Media & Library Sciences 42 (3) (2005) 313–328.
[18] Y.C. Shiue, G.M. Guo, A method for building core journal ranked lists in electronic commerce subject area, International Journal of Electronic Business Management 3 (2) (2005) 151–169.
[19] Y.C. Shiue, R.Y. Chang, G.M. Guo, A computer-aided bibliometrics system for journal citation analysis and departmental core journal ranking list generation, Journal of Educational Media & Library Sciences 42 (2) (2004) 199–220.