The Evolution of core collections can be described via Banach space valued stochastic processes

DOI: 10.1016/S0895-7177(98)00141-1
Handle: http://hdl.handle.net/1942/798

THE EVOLUTION OF CORE COLLECTIONS CAN BE DESCRIBED VIA BANACH SPACE VALUED STOCHASTIC PROCESSES

by

L. Egghe

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, 2610 Wilrijk, Belgium

ABSTRACT

Core collections, as e.g. the set of source journals, selected by the Institute for Scientific Information, vary from year to year : most of them stay, some leave and others enter.

In this paper we show that the mechanism of this process can be revealed by using stochastic processes with values in an infinite dimensional Banach space. More concretely we show that the evolution can be described by a quasi-martingale with values in the Hilbert space L^2. Criteria for the convergence of these processes are given, leading to a stable limit set of core items, for t (time) high.

Keywords : source journal, core collection, Banach space, stochastic process, quasi-martingale.

[1]Permanent address

The author is indebted to Prof. Dr. R. Rousseau for interesting discussions on the topic of this paper.

I. INTRODUCTION

Core collections can be defined in any context where one considers a basic set of items w.r.t. a certain usage, e.g. a collection of (basic) algebra journals. The most famous example is the set of source journals (also called core journals) as defined, on a yearly basis, by the Institute for Scientific Information (ISI). This set can be found e.g. in the yearly editions of the Journal Citation Reports (JCR), a statistical side product of the ISI data. Core collections can even be considered outside the journal scope. An example is offered by a core collection of books in a library (where the core could be defined according to the number of times the books are checked out, or via another criterium).

It is hence clear that

- core collections vary from year to year

and that

- their evolution in time is an important informetric isue.

Indeed, going back to the example of source journals of ISI, the fact that a journal is or is not a source journal has many important consequences. Having a publication in a source journal means that the described research is visible in the world. The researcher can mention this in the annual report of the research group which has consequences for the financial support of this group. This is certainly the case in most Belgian universities (including mine) and I am sure that this is also applying to other research centers in the world. Being a source journal also means that this journal is part from a whole mechanism of evaluations (based on the degree of citedness) performed by ISI but also by other research institutes. Obviously these journals are the ones with an impact factor (IF). This measure alone, when related to the specific field, indicates the degree of research-orientedness of the journal. The set of source journals in a certain scientific field determine the way this field is studied (obsolescence, communication networks between authors or between journals, and so on).

This is not the place to review the vast literature on citation analysis. In fact, as described above, our scope is much wider. Let us content ourselves by referring to [1], chapter III for a 90-page account on the uses of citation analysis.

The problem of determining a core collection and of studying its evolution in time has been noticed in [2]. Of course, much earlier, one can read discussions on the selection of the core collection set, especially in the case of source journals, see e.g.[3-9]. But in the author's opinion the paper of Rousseau and Spinak is the first to mention the several methodological questions around core collections.

The basic problem can be formulated as follows. Suppose at a certain time (say $t=1$) we start with a core collection $A=A_1$. How can we determine the core collections A_t at later times t ? Is the process stable in the sense that the sets A_t converge in some way ?

In [10] we studied a heavily simplified version of this problem. There we started with a set A and studied, at every t, what is left from this set. Such results only indicate the quality of the original set of source journals.

Going back to the basic problem, we are requested to indicate the probabilities with which we can expect a certain set (subset of the universe U under consideration) to be the set of core items A_t. In the next section we will investigate this by using, for every set A, its characteristic function

$$\chi_A : U \to \{0,1\}$$

$$u \to \chi_A(u) \begin{cases} =1 \text{ if } u \in A \\ =0 \text{ if } u \notin A \end{cases}$$

This function indeed characterises the set A.

We will be using stochastic processes. Although these probabilistic items are not well-known in our field we do not introduce them here ; for this we refer to [10,11] where extensive appendices are devoted to an explanation of these important tools.

So, the next section will introduce the stochastic process that deals with the evolution of the core collection. We will also investigate its convergence properties, i.e. we will indicate whether or not (and when) a stable limit core collection exists, for t high.

The last section gives some remarks and open problems.

II. THE STOCHASTIC PROCESS THAT DESCRIBES THE EVOLUTION OF CORE COLLECTIONS.

II.1. Introduction

Let U be our universe of all items that are considered for becoming a core item. As an example, U can be the total collection of journals (possibly restricted to a field under study). U can contain journals that will only exist from a certain time on ; before that time, obviously, their chance to become a source journal is zero. In the same way it will contain journals that stop after a certain time. These journals will then have a zero change to be a source journal, after that time. The same can be said in the case of core books in a library w.r.t. purchase of new books or the weeding of old books. In the sequel we will henceforth use the terminology : U is the set of all items and we study the evolution of core items from U.

Suppose, at the starting point $t=1$, we have $A=A_1 \subset U$ as the core set. Instead of allowing t to be a fixed time period, we will consider t as the time in which t-1 elementary changes in the core set have been executed, i.e. one core item out or one non-core item in. More concretely we will move from $A=A_1$ to A_2 as the result of

- either one element from A is deleted as a core item, or,
- one element from $U \backslash A$ enters as a core item.

In general, and denoting by A_t the set of core items at time t, we move from t to $t+1$ if one of the following actions have been taken place :

- either one element from A_t is deleted as a core item, or,
- one element from $U\backslash A_t$ enters as a core item.

In the former case, and denoting by $u\in A_t$ this element, we have that $A_{t+1}=A_t\backslash\{u\}$. In the latter case, and denoting by $u\in U\backslash A_t$ this element, we have that $A_{t+1}=A_t\cup\{u\}$. This are the only two alternatives, of course applicable to every $u\in U$.

As said in the previous section, we will study the evolution of A_t by means of its characteristic function χ_{A_t}, indeed characterizing A_t. As such, χ_{A_t} is not only a function of $u\in U$ but also of the probability space $(\Omega,\mathcal{F},P)$ that we are about to construct. Hence every χ_{A_t} is a random variable (r.v.) with values in some function space (to be determined in the sequel).

II.2 Construction of the stochastic process.

Let us suppose we are at time t and that $A_t \subset U$ is the core set at t. We will indicate how we will pass from t to t+1. For each $u\in A_t$ we put

$$\alpha(t)(u) = P(u\notin A_{t+1} \,|\, u\in A_t) \tag{1}$$

and for each $u\in U\backslash A_t$ we put

$$\beta(t)(u) = P(u\in A_{t+1} \,|\, u\in U\backslash A_t) \tag{2}$$

These are the only #U (# = number of elements) cases leading to a change in the core set, hence bringing us from t to t+1. By the very definition of t+1 we have

$$\sum_{u\in A_t} \alpha(t)(u) + \sum_{u\in U\backslash A_t} \beta(t)(u) = 1 \tag{3}$$

The underlying probability spaces and σ-algebras are contructed as follows. For each t, define

$$\Omega_t = \prod_{i=1}^{t} U_i \tag{4}$$

where $U_i=U$ for every i. Hence Ω_t is the t-fold product of identical copies of U. Furthermore we consider the points of Ω_t as subsets of

$$\Omega = \prod_{i=1}^{\infty} U_i \tag{5}$$

by identifying, for each t, $(u_1,\ldots,u_t)\in\Omega_t$ with

$$\prod_{i=1}^{t} \{u_i\} \times \prod_{j=t+1}^{\infty} U_j \tag{6}$$

By this identification we have that, if $\mathcal{F}_t=\mathcal{P}(\Omega_t)$, the set of all subsets of Ω_t, the sequence $(\mathcal{F}_t)_{t\in N}$ is an increasing sequence of σ-algebras. Denote by $\mathcal{F}$ the σ-algebras generated by $\bigcup_{t=1}^{\infty}\mathcal{F}_t$. The probabilities are constructed as follows : at t, for every $(u_1,\ldots,u_t)\in\Omega_t$ and $u\in U$,

$$P_{t+1}(u_1,\ldots,u_t,u) = P_t(u_1,\ldots,u_t).\alpha(t)(u) \tag{7}$$

if $u\in A_t$ and

$$P_{t+1}(u_1,\ldots,u_t,u) = P_t(u_1,\ldots,u_t).\beta(t)(u) \tag{8}$$

if $u\in U\backslash A_t$ (always using identification (6)). At each level, α(t)(u) and β(t)(u) are $\mathcal{F}_t$-measurable functions on Ω_t.

Formulae (7) and (8) determine, inductively, the probability measures for all $t\geq 1$, where, if $t=1$, we define $P_1(u)=1$ if $u\in A_1$ and $P_1(u)=0$ if $u\in U\backslash A_1$. Let us also denote by P the product probability measure on Ω, determined by the probability measures P_t on Ω_t (again using the said identification). Because of this identification we can us the notation P instead of P_t on Ω_t. On the system $(\Omega,\mathcal{F}_t,P)$ we can define the following stochastic process : since, in case (1) applies, χ_{A_t} is changed into $\chi_{A_t\backslash\{u\}}$ and in case (2) applies, χ_{A_t} is changed into $\chi_{A_t\cup\{u\}}$ we have the following conditional expectation equation for $X_{t+1} = \chi_{A_{t+1}}$:

$$E^{\mathcal{F}_t}(\chi_{A_{t+1}}) = \sum_{u\in A_t} \alpha(t)(u)\chi_{A_t\setminus\{u\}} + \sum_{u\in U\setminus A_t} \beta(t)(u)\chi_{A_t\cup\{u\}}\,, \qquad (9)$$

by the law of total change.

For all functions we have dropped the $u\in U$ and $\omega\in\Omega$-dependency notation. In full notation we would have (e.g. for χ_{A_t})

$$\chi_{A_t(\omega)}(u)\,,$$

$\omega\in\Omega_t$, $u\in U$ and where the function $\omega\to\chi_{A_t(\omega)}$ is $\mathcal{F}_t$-measurable with values in some function space (to be determined later). It is this function (but for $t+1$) that is used in the $E^{\mathcal{F}_t}(.)$ in formula (9).

From (9) we derive (see Appendix A) :

$$E^{\mathcal{F}_t}(\chi_{A_{t+1}}) = (1-\alpha(t)-\beta(t))\chi_{A_t}+\beta(t) \qquad (10)$$

where $\alpha(t)$ and $\beta(t)$ denote the functions $u\to\alpha(t)(u)$ and $u\to\beta(t)(u)$, $u\in U$. Since, for every $u\in U$, $\alpha(t)(u)$ and $\beta(t)(u)$ are $\mathcal{F}_t$-measurable real functions on Ω_t we have that $\alpha(t)$, $\beta(t)$ are $\mathcal{F}_t$-measurable function-valued functions on Ω_t. To be more concrete we suppose that all these functions take values in $L^2(U,\mathcal{G},\mu)$, where $(U,\mathcal{G},\mu)$ is a probability space on U and where L^2 denotes the set of functions which squares are integrable. Note that all functions χ_{A_t} above belong to $L^2(U,\mathcal{G},\mu)$. $(U,\mathcal{G},\mu)$ can be anything, even (in finite cases) the counting measure space (see e.g. [12]).

This completes the construction of the Hilbert-space valued process $(X_t,\mathcal{F}_t,P)$ where $X_t=\chi_{A_t}$. This process gives a complete description of the evolution of the set of core items A_t when t passes, once the probabilities of transition are known.

We will now investigate the properties of this process.

II.3 Properties of $(X_t,\mathcal{F}_t,P)$.

From (10) it is clear that

$$E^{\mathcal{F}_t}X_{t+1}-X_t = -(\alpha(t)+\beta(t))X_t+\beta(t) \tag{11}$$

Hence ($\|.\|_2$ denotes the L^2-norm)

$$E(\|E^{\mathcal{F}_t}(X_{t+1})-X_t\|_2)$$

$$= E(\|\beta(t)-(\alpha(t)+\beta(t))X_t\|_2)$$

$$= E(\|\beta(t)(1-X_t)-\alpha(t)X_t\|_2)$$

$$\leq E(\|\beta(t)(1-X_t)\|_2)+E(\|\alpha(t)X_t\|_2)$$

$$\leq E(\|\beta(t)\|_2)+E(\|\alpha(t)\|_2) \tag{12}$$

Since X_t takes values in $\{0,1\}$. Hence if we suppose

$$\sum_{t=1}^{\infty} E(\|\alpha(t)\|_2) < \infty \tag{13}$$

$$\sum_{t=1}^{\infty} E(\|\beta(t)\|_2) < \infty \tag{14}$$

then we have that $(X_t,\mathcal{F}_t)$ is a quasi-martingale (cf. [13,14]). Since $(X_t)_{t\in N}$ is $\|.\|_2$-uniformly bounded (by 1) they converge a.e. to an integrable function X_∞, i.e. an integrable function on $(\Omega,\mathcal{F},P)$ with values in $L^2(U,\mathcal{G},\mu)$, if this Banach space has the Radon-Nikodym-Property (RNP) (see [13,14]). This is so since L^2 is a Hilbert space, see [15]. In this book one can find background information on (RNP) Banach spaces. This are spaces in which the "classical" theorem of Radon-Nikodym (sometimes called Lebesgue-Radon-Nikodym) is valid, also ensuring a.e. convergence of processes as above.

Suppose then that (13) and (14) are true. We hence have the existence of a function

$$X_\infty \in L^1_{L^2}(\Omega,\mathcal{F},P) \tag{15}$$

the space of integrable functions on $(\Omega,\mathcal{F},P)$ with values in $L^2=L^2(U,\mathcal{G},\mu)$ such that

$$\|.\|_2\text{-}\lim_{t\to\infty} X_t = X_\infty \ , \ P\text{-a.e.} \tag{16}$$

This means that

$$\lim_{t\to\infty} \int_U \|X_t(\omega)-X_\infty(\omega)\|^2 = 0 \tag{17}$$

for $\omega\in\Omega\backslash B$, where $P(B)=0$. For every $\omega\in\Omega\backslash B$ we hence have an L^2-convergent sequence $(X_t(\omega))_{t\in\mathbb{N}}$, hence L^1-convergent (this is so on finite measure spaces as $(U,\mathcal{G},\mu)$). Hence there is an a.e. convergent subsequence. For these results on measure theory, see [12]. Hence there exists $(t_n)_{n\in\mathbb{N}}$, a strictly increasing sequence in $\mathbb{N}$ (dependent on ω!) such that

$$\lim_{n\to\infty} X_{t_n}(\omega) = X_\infty(\omega) \tag{18}$$

Since $X_{t_n}(\omega)=\chi_{A_{t_n}(\omega)}$, the values of $X_\infty(\omega)$ can only be 0 or 1. This is trivial. Define then

$$A_\infty = \{\omega\in\Omega\|X_\infty(\omega)=1\} \tag{19}$$

Then A_∞ is our stable limit set since, obviously,

$$X_\infty = \chi_{A_\infty} \tag{20}$$

and by (16) or (17).

We hence have proved the following theorem.

Theorem : Let the process $(X_t,\mathcal{F}_t,P)$ describe the evolution of the set of core items as defined by formula (9). Suppose that (13) and (14) are valid. Then the process is a uniformly bounded quasi-martingale in $L^2(U,\mathcal{G},P)$ which $\|.\|_2$-converges a.e. to an integrable function. This function determines uniquely a stable limit set which is the limiting set of core items.

III. REMARKS AND OPEN PROBLEMS.

III. 1 Remarks

We could also study the probability that an item $u\in U$ becomes a core item at time t :

$$P(u,t) = P \text{ (u is core item at t)}$$

By the law of total change we have

$$\begin{aligned} P(u,t) &= P \text{ (u is core item at t)} \\ &= P \text{ (u is core item at t } | \text{ u is core item at t-1)} \\ &\qquad .P \text{ (u is core item at t-1)} \\ &\quad + P \text{ (u is core item at t } | \text{ u is not a core item at t-1)} \\ &\qquad .P \text{ (u is not a core item at t-1)} \\ &= (1-a(t-1,u))P(u,t-1)+b(t-1,u)(1-P(u,t-1)), \end{aligned} \tag{21}$$

where (for all $t\geq 1$)

$$a(t,u) = P(\text{u is not a core item at t } | \text{ u is core item at t-1})$$

and

$$b(t,u) = P(\text{u is core item at t } | \text{ u is not a core item at t-1}).$$

Hence

$$P(u,t) = (1-a(t-1,u)-b(t-1,u))P(u,t-1)+b(t-1,u) \qquad (22)$$

Formula (22) is comparable with formula (10) but for probabilities instead of random variables.

Formula (22) gives the overall probability for $u \in U$ to be a core item at t, unconditionally. From (22) we deduce inductively

$$P(u,t) = \prod_{i=1}^{t-1}(1-a_i-b_i)\frac{\#A}{\#U} + \prod_{i=2}^{t-1}(1-a_i-b_i)b_1 + \prod_{i=3}^{t-1}(1-a_i-b_i)b_2 + \ldots + (1-a_{t-1}-b_{t-1})b_{t-2} \qquad (23)$$

, where $\frac{\#A}{\#U} = P(u,1)$, since A is the (given) core set at $t=1$.

III.2 Problems.

1. Say we start, at $t=1$, with another core set $B \subset U$ (instead of A). Determine the limit set B_∞ in this case and compare it with the one we found above : A_∞. This problem, dealing with the stability of the system was formulated in [2]. An application of this would be an understanding of the evolution of e.g. a set of source journals constituted by another institute than ISI (possibly from another country, thereby possibly focusing other (more local) journals : we could then see if, for large t, we recover (more or less) the source journals as defined by ISI.
2. How can rankings (e.g. based on impact factors) be involved in these models? What is the relation of a ranking (at a certain time t) of a source journal with the number of times the journal was a source journal?
3. What is the stability of the results in the sense that small changes to the values of α and β should result in small changes in the final result.
4. Apply these models to other domains in information science : core collections of books in a library, evolution of retrieved sets of documents over time (w.r.t. a fixed query), evolutions of bibliographies, of research groups, etc.

References

[1] L. Egghe and R. Rousseau, *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science*. Elsevier, Amsterdam (1990).

[2] R. Rousseau and E. Spinak, Do a field list of internationally visible journals and their journal impact factors depend on the initial set of journals ? A research proposal. *Journal of Documentation* **52**(4), 449-456 (1996).

[3] L. Velho, The "meaning" of citation in the context of a scientific peripheral country. *Scientometrics* **9**, 71-89 (1986).

[4] L. Velho, The author and the beholder : how paradigm commitments can influence the interpretation of research results. *Scientometrics* **11**, 59-70 (1987).

[5] J. Gaillard, La science du tiers monde est-elle visible ? *La Recherche* **20**, 636-640 (1989).

[6] P. Vinkler, Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics* **10**, 157-177 (1986).

[7] E. Garfield, Journal citation studies 32. Canadian journals, Part 2 : an analysis of Canadian research published at home and abroad. *Current Contents*, August 20. Reprinted in : *Essays of an Information Scientist* **4** (ISI, Philadelphia), 249-253 (1981).

[8] B. Cronin, Transatlantic citation patterns in educational psychology. *Education Libraries Bulletin* **24**, 48-51 (1981).

[9] H. Inhaber and M. Alvo, World science as an input-output system. *Scientometrics* **1**, 43-64 (1978).

[10] L. Egghe, Dynamics of a field list of internationally visible journals : a stochastic model. *Preprint* (1998).

[11] L. Egghe and R. Rousseau, Stochastic processes determined by a general success-breeds-success principle. *Mathematical and Computer Modelling* **23**(4), 93-104 (1996).

[12] P.R. Halmos, *Measure Theory*. Graduate texts in Mathematics **18**, Springer Verlag, New York (1974).

[13] L. Egghe, *Stopping Time Techniques for Analysts and Probabilists*. London Mathematical Society Lecture Notes Series **100**. Cambridge University Press, Cambridge, UK (1984).

[14] G.A. Edgar and L. Sucheston, *Stopping Times and directed Processes*. Cambridge University Press, Cambridge, UK (1992).

[15] J. Diestel and J.J.Jr.Uhl, *Vector Measures*. American Mathematical Society Surveys **15**, AMS, Providence, RI, USA (1977).

Appendix A. Proof of formula (10)

Formula (9) reads

$$E^{\mathcal{F}_t}(\chi_{A_{t+1}}) = \sum_{u\in A_t} \alpha(t)(u)\chi_{A_t\setminus\{u\}} + \sum_{u\in U\setminus A_t} \beta(t)(u)\chi_{A_t\cup\{u\}}$$

$$= \sum_{u\in A_t} \alpha(t)(u)\chi_{A_t} - \sum_{u\in A_t} \alpha(t)(u)\chi_{\{u\}}$$

$$+ \sum_{u\in U\setminus A_t} \beta(t)(u)\chi_{A_t} + \sum_{u\in U\setminus A_t} \beta(t)(u)\chi_{\{u\}}$$

So, for $u'\in A_t$ we have

$$(E^{\mathcal{F}_t}(\chi_{A_{t+1}}))(u')$$

$$= \sum_{u\in A_t} \alpha(t)(u) - \alpha(t)(u')$$

$$+ \sum_{u\in U\setminus A_t} \beta(t)(u)$$

$$= 1 - \alpha(t)(u'))$$

and for $u' \in U \setminus A_t$ we have

$$(E^{\mathcal{F}_t}(\chi_{A_{t+1}}))(u') = \beta(t)(u')$$

Hence, $E^{\mathcal{F}_t}(\chi_{A_{t+1}})$ is the L^2-function

$$\begin{aligned} E^{\mathcal{F}_t}(\chi_{A_{t+1}}) &= (1-\alpha(t))\chi_{A_t} + \beta(t)\chi_{U\setminus A_t} \\ &= (1-\alpha(t))\chi_{A_t} + \beta(t)(1-\chi_{A_t}) \\ E^{\mathcal{F}_t}(\chi_{A_{t+1}}) &= (1-\alpha(t)-\beta(t))\chi_{A_t} + \beta(t), \end{aligned}$$

proving (10).