

NHỮNG THÔNG TIN CƠ BẢN VỀ TRẮC LƯỢNG THƯ MỤC

1. Lịch sử hình thành và quá trình phát triển (Nguyễn Hữu Viêm)

Thuật ngữ Trắc lượng thư mục (Bibliometrics) xuất hiện lần đầu tiên trên “Tập chí Tư liệu” (Journal of Documentation) số tháng 12 năm 1969, trong bài báo của Alan Pritchard, với nhan đề: “Thư mục thống kê hay Trắc lượng thư mục?” (Statistical Bibliography or Bibliometrics?). Pritchard cho rằng, thuật ngữ Thư mục thống kê phải được thay bằng một thuật ngữ chính xác hơn, phản ánh đúng hơn nội dung. Ông viết: “Thuật ngữ Statistical Bibliography không rõ nghĩa, có thể làm người ta dễ lẫn lộn với chính bản thân Thống kê học hoặc Thư mục về thống kê học”. Ông đưa ra một thuật ngữ mới là Bibliometrics (Trắc lượng thư mục), có nghĩa là áp dụng toán học và các phương pháp thống kê đối với sách và các phương tiện truyền thông khác.

Từ đó, thuật ngữ mới do Alan Pritchard nêu ra đã được giới chuyên môn chấp nhận. Đôi khi người ta cũng sử dụng thuật ngữ Statistical Analysis (Phân tích thống kê) như một thuật ngữ tương đương.

Cũng vào năm 1969, các nhà khoa học Liên Xô đưa ra thuật ngữ Scientometrics (Trắc lượng khoa học), với quan niệm gần như quan niệm của A. Pritchard. Năm 1988, các nhà khoa học Viện Thông tin KHKT toàn Liên bang (VINITI), Liên Xô đưa ra thuật ngữ Informetrics (Trắc lượng thông tin), sử dụng trong lĩnh vực thông tin khoa học.

Từ đó đã hình thành mối quan hệ giữa hai thuật ngữ Thư mục thống kê đã có trước, và Trắc lượng thư mục mới ra đời. Do đó,

người ta coi lịch sử Trắc lượng thư mục được bắt đầu từ khi kết hợp thư mục học và thống kê học thành Statistical Bibliography, vào khoảng những năm 20 thế kỷ XX.

Thuật ngữ Trắc lượng thư mục sau khi được chấp nhận (trong khoảng 40 năm qua) đã có bước phát triển mới, từ một phương pháp nghiên cứu định tính trở thành một lĩnh vực được các nhà chuyên môn Bibliometrician (Nhà nghiên cứu Trắc lượng thư mục) đào sâu nghiên cứu.

Trắc lượng thư mục khai thác phân tích số lượng và thống kê đối với các yếu tố miêu tả chính trong thư mục học trong một lĩnh vực nhất định. Các bản thư mục và CSDL thư mục của các cơ quan khoa học đều có thể là đối tượng khảo sát, nghiên cứu phân tích thống kê. Các nhà nghiên cứu có thể sử dụng Trắc lượng thư mục để đánh giá và xác định mức độ ảnh hưởng của một tác giả hoặc mô tả mối quan hệ qua lại giữa hai hoặc nhiều tác giả, hoặc tác phẩm. Thông thường trong nghiên cứu Trắc lượng thư mục, người ta hay sử dụng các CSDL Social Science Citation Index (SSCI- Chỉ dẫn trích dẫn khoa học xã hội), Science Citation Index (SCI- Chỉ dẫn trích dẫn khoa học) hoặc Arts and Humanities Citation Index (AHCI- Chỉ dẫn trích dẫn nghệ thuật và khoa học nhân văn), do Viện Thông tin Khoa học Philadelphia, Hoa Kỳ (Institute for Scientific Information-ISI) xây dựng.

Gần đây phương pháp thống kê được vận dụng để nghiên cứu mức độ ảnh hưởng qua lại và tác động của những trang Web đến các trang Web khác trên Internet dựa trên số lần chúng được liên kết (hypelink) tới các trang Web khác và được gọi là

Nghiên cứu - Trao đổi

Webometrics hoặc Cybematics (trắc lượng mạng).

Trong một công trình của Wolfgang Glanzl xuất bản năm 2003 [2], Trắc lượng thư mục được xác định gồm ba hợp phần:

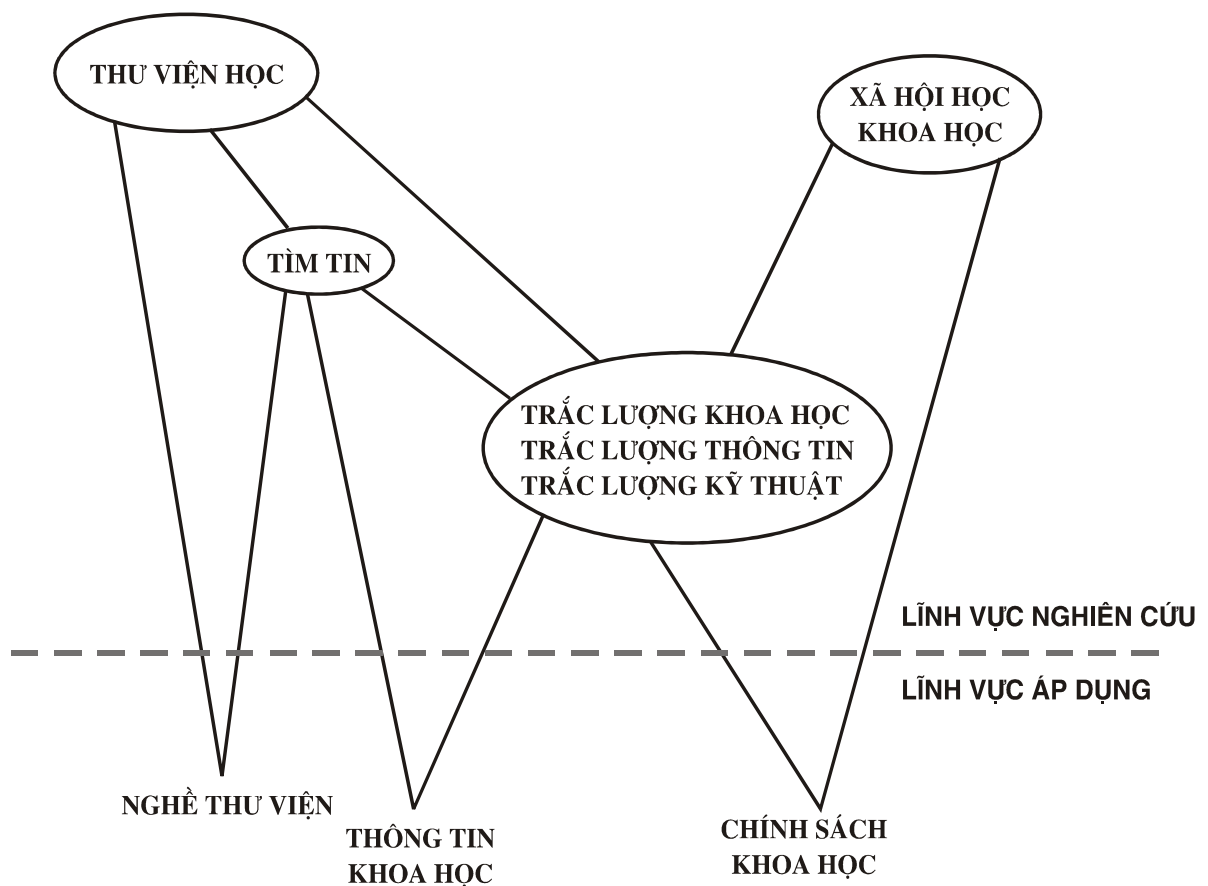
- *Trắc lượng thư mục của các nhà trắc lượng thư mục* (phương pháp luận). Đây là khu vực nghiên cứu trắc lượng thư mục cơ bản. Nghiên cứu phương pháp luận trong trắc lượng thư mục được thực hiện chủ yếu ở khu vực này.

- *Trắc lượng thư mục của các lĩnh vực khoa học* (thông tin khoa học). Các nhà nghiên cứu trắc lượng thư mục trong các lĩnh vực khoa học cụ thể gồm nhiều nhóm

rất đa dạng liên quan chặt chẽ với chuyên ngành của họ. Khu vực này được xem là phần mở rộng của thông tin khoa học bằng các phương pháp trắc lượng.

- *Trắc lượng thư mục của chính sách và quản lý khoa học* (chính sách khoa học). Đây là khu vực đánh giá nghiên cứu những vấn đề quan trọng nhất trong trắc lượng thư mục. Cấu trúc khoa học của từng quốc gia, vùng và cơ quan khoa học, so sánh giữa chúng thường được thực hiện ở khu vực này.

W. Glanzl cũng đã thể hiện mối quan hệ của Trắc lượng thư mục với các lĩnh vực khoa học khác trong một sơ đồ như sau:



Gần 20 năm sau khi thuật ngữ Bibliometrics ra đời, Hội nghị quốc tế đầu tiên về Bibliometrics được triệu tập tại Bỉ vào tháng 8 năm 1987, sau đó 2 năm một lần, đến nay đã được 11 kỳ. Năm 2007, Hội nghị quốc tế kỳ 11 được triệu tập tại Madrid, Tây Ban Nha. Các cuộc Hội nghị quốc tế được tổ chức tại nhiều vùng khác nhau trên thế giới, tại châu Á đã tổ chức 2 kỳ, kỳ thứ 3 tại Bangalore, Ấn Độ tháng 8 năm 1991 và kỳ thứ 9 được tổ chức tại Bắc Kinh, Trung Quốc tháng 8 năm 2003.

Tại Hội nghị quốc tế kỳ thứ 6 tại Jerusalem, Israel vào tháng 7 năm 1997, đã quyết định xuất bản Tạp chí quốc tế về Trắc lượng thư mục, Trắc lượng khoa học và Trắc lượng thông tin (The International Journal of Bibliometrics, Scientometrics and Informetrics). Tạp chí này có thể đọc trên Internet với tên Cybermetrics (Trắc lượng mạng).

2. Các định luật cơ bản trong trắc lượng thư mục (ThS. Trần Mạnh Tuấn)

Thành tựu lớn đã được ghi nhận của phương pháp đo lường thư mục là việc tìm ra 3 định luật mang tên nhà khoa học đã phát hiện ra: Định luật Lotka (Alfred James Lotka) về sáng tạo khoa học của cá nhân nhà khoa học, Định luật Bradford (Samuel C. Bradford) về sự tản mạn tin trong các tạp chí khoa học, và Định luật Zipf (George Kinsley Zipf) về quan hệ giữa tần suất xuất hiện các từ và thứ hạng của chúng trong văn bản¹. Dưới đây là nội dung chính của những định luật này.

Định luật Lotka phản ánh mối quan hệ

giữa sức sáng tạo/mức độ đóng góp của cá nhân nhà khoa học với số lượng các nhà khoa học; sự phân bố số lượng công trình khoa học trong đội ngũ các nhà khoa học trong một lĩnh vực/ngành khoa học xác định. Dưới đây là nội dung chính của định luật.

Số lượng tác giả tạo nên các công trình khoa học là một hàm số của số lượng các công trình khoa học được chính họ tạo nên và được xác định theo công thức tổng quát:

$$N_n = 1/n^2 \times 60 \% \times N_{(s)}$$

Trong đó:

- N_n : Số lượng các tác giả công bố n công trình khoa học,
- n: Số lượng công trình khoa học được công bố của tác giả đang được khảo sát,
- $N_{(s)}$: Tổng số các nhà khoa học trong lĩnh vực đang được khảo sát.

Theo công thức tổng quát trên, trong một lĩnh vực khoa học xác định, có khoảng 60% nhà khoa học thuộc lĩnh vực đó chỉ công bố 1 công trình khoa học, 15% công bố 2 công trình khoa học ($1/2^2 \times 60\% \times N_{(s)}$) và 7% công bố 3 công trình khoa học ($1/3^2 \times 60\% \times N_{(s)}$), ...

Theo thống kê của A.J. Lotka về khả năng sáng tạo của nhà khoa học, chỉ có 6% các tác giả trong một lĩnh vực khoa học cụ thể sẽ công bố trên 10 công trình khoa học. Nhận xét này mang tính chất thống kê, các giá trị nhận được chỉ chính xác khi thỏa mãn đủ các điều kiện xác định (ví dụ khoảng thời gian được khảo sát đủ lớn,...). Định luật Lotka được sử dụng để xác định

¹ Các thông tin, số liệu được trình bày trong phần giới thiệu 3 định luật này được tập hợp từ các nguồn khác nhau, trong đó có [1]; [2]; [3] trong phần Tài liệu tham khảo.

mức đóng góp của mỗi cá nhân khoa học.

Định luật Lotka mang đậm yếu tố lịch sử và thực tiễn tại thời điểm được tìm ra cách đây gần một thế kỷ (1926). Trong khoảng thời gian này, sự phát triển các ngành khoa học, nguồn tài liệu khoa học và năng lực sáng tạo của nhà khoa học đã có những thay đổi mang tính đột biến dưới sự tác động của rất nhiều yếu tố khác nhau. Giá trị của Định luật chính là ở chỗ A.J. Lotka đã xác nhận *sự phụ thuộc* của năng lực sáng tạo của nhà khoa học vào số lượng các nhà khoa học - cộng đồng mà nhà khoa học đó là một thành viên - của lĩnh vực đó. Ngày nay, cùng với khả năng của con người nói chung trong xã hội là vượt tới các giới hạn *nhANH hơn, xa hơn, cao hơn, mạnh hơn* năng lực sáng tạo của nhà khoa học ngày càng được cải thiện một cách rõ rệt nhờ sự trợ giúp của *nhIEU loại phương tiện* trong các quá trình trao đổi và thu nhận thông tin; trợ giúp họ kiểm soát được thông tin có liên quan ngày một hiệu quả hơn trong bối cảnh của sự bùng nổ thông tin; trợ giúp họ trong việc giải những bài toán công kênh, phức tạp song có thể mô phỏng bằng các thuật toán (vấn đề xử lý thông tin nói chung);.... Như vậy, nguồn di sản trí tuệ tinh thần của loài người được tăng trưởng với một gia tốc ngày càng cao, số lượng các nhà khoa học ngày càng lớn, và vì thế, công thức mà A.J. Lotka xây dựng trước kia cần phải được bổ sung nhiều tham số phức tạp khác. Hơn nữa, có thể dự báo mối quan hệ giữa sức sáng tạo của nhà khoa học với

quy mô của đội ngũ nhà khoa học sẽ không còn mang tính chất tuyến tính như thuở ban đầu nó được khám phá. Rất tiếc cho đến nay, các hướng nghiên cứu mà A.J. Lotka theo đuổi vẫn chưa được tiếp tục. Có lẽ, các chỉ số thống kê dựa trên một *Định luật Lotka mới*, nếu có, chắc chắn cùng với các số liệu thống kê về các công trình đăng trên *Danh mục tạp chí khoa học Philadelphia*² sẽ trở thành các công cụ hết sức đặc sắc để đánh giá mức độ đóng góp và trình độ phát triển khoa học một cách hoàn chỉnh hơn.

Định luật Bradford được sử dụng để xác định số lượng các tạp chí hạt nhân trong một lĩnh vực/ngành khoa học cụ thể. Nội dung định luật:

Danh mục các tạp chí trong một lĩnh vực khoa học cụ thể có thể được chia thành 3 phần và được phân chia sao cho mỗi phần là danh sách các tạp chí mà trên đó đăng tải cùng một số lượng các bài nghiên cứu trong lĩnh vực đó.

Một bộ phận danh mục tạp chí đặc biệt, nơi mật độ bài nghiên cứu rất lớn so với các nhóm khác của lĩnh vực này, công bố khoảng 1/3 số bài báo, chiếm số lượng rất ít trong tổng số các tên tạp chí được khảo sát, được gọi là danh mục tạp chí hạt nhân - phần danh mục các tên tạp chí cốt lõi của lĩnh vực đang khảo sát.

Các tạp chí ở phần danh mục thứ hai cũng chứa khoảng 1/3 số bài nghiên cứu như ở phần hạt nhân, song số lượng các tên tạp chí của danh mục này lại lớn hơn nhiều so với số lượng tên tạp chí trong

² Một danh mục bao gồm tạp chí có uy tín trên thế giới do Viện Thông tin Khoa học (ISI) tại Philadelphia (Mỹ) xây dựng trên cơ sở thống kê các trích dẫn được thực hiện. Tuy rất có giá trị song nhiều chuyên gia cho rằng, phương pháp đánh giá này còn bộc lộ nhiều hạn chế và thiếu toàn diện.

danh mục hạt nhân.

Ở phần danh mục thứ 3, cũng chứa khoảng 1/3 số bài tạp chí còn lại, song số lượng tên tạp chí lại lớn hơn rất nhiều so với ngay cả số lượng tên tạp chí trong danh mục thuộc nhóm thứ hai nêu trên.

Trên cơ sở khảo sát thống kê về sự phân bố các bài tạp chí thuộc lĩnh vực *Địa vật lý* (Geophysics), bao gồm 326 tên tạp chí, S.C. Bradford đã có được các số liệu thống kê sau:

- Phần danh mục tạp chí hạt nhân-nhóm 1- gồm 9 tên tạp chí, chứa 429 bài nghiên cứu.

- Phần danh mục tạp chí nhóm 2 gồm 59 tên tạp chí, chứa 499 bài nghiên cứu.

- Phần danh mục tạp chí nhóm 3 gồm 258 tên tạp chí, chứa 404 bài nghiên cứu.

Từ các số liệu thống kê trên, S.C. Bradford đã xây dựng tỷ lệ phân bố số tên tạp chí theo 3 bộ phận phân chia như trên là $1 : n : n^2$.

Như vậy, số tên tạp chí ở nhóm hạt nhân là 9; ở nhóm 2 là 59 (vào khoảng 9×5^1); và ở nhóm 3 là 258 (ước khoảng 9×5^2). Trong khi đó, mỗi nhóm các tên tạp chí trên đều có khoảng trên 400 bài viết.

Định luật Bradford được ứng dụng chủ yếu trong việc:

- Hoạch định chính sách bổ sung nguồn tin (là các tạp chí khoa học) đối với một cơ quan, đặc biệt là đối với một hệ thống các cơ quan thông tin-thư viện liên kết với nhau theo mô hình một consortium bổ sung; xác định danh mục các tạp chí khoa học thiết yếu để tạo lập một sản phẩm thông tin nào đó, ví dụ tạp chí tóm tắt (vấn

đề được quan tâm chủ yếu trong thực tiễn công tác thông tin-thư viện);

- Nghiên cứu, khảo sát quy luật phát triển nguồn tạp chí khoa học trong quan hệ với sự phát triển các ngành, lĩnh vực khoa học (vấn đề được quan tâm chủ yếu trong thông tin học);

- Nghiên cứu, khảo sát quá trình tích hợp giữa các ngành khoa học thông qua việc nghiên cứu, khảo sát quá trình tích hợp các nguồn thông tin khoa học được tạo nên (vấn đề được quan tâm chủ yếu trong khoa học luận)....

Như đã biết, tạp chí là một loại ấn phẩm khoa học đặc biệt, xuất hiện từ đầu thế kỷ XVII, và là phương tiện công bố *nhANH NHẤT* các ý tưởng và khám phá khoa học *MỚI NHẤT*. Bởi vậy, hệ thống tạp chí khoa học nói chung sẽ phản ánh một cách đầy đủ nhất *trình độ, quy mô, các trào lưu và khuynh hướng* phát triển khoa học. Ngày nay, nhịp độ gia tăng lượng tạp chí khoa học ước khoảng 12-15%/năm, xu hướng giao thoa, tích hợp giữa các ngành, lĩnh vực khoa học được diễn ra ngày càng phổ biến và sâu sắc. Chính vì thế, ý nghĩa thực tiễn của Định luật Bradford đối với các nghiên cứu về sự phát triển các ngành khoa học, đối với việc xây dựng chiến lược và các chính sách phát triển nguồn thông tin khoa học của các cơ quan thông tin, thư viện, đối với chiến lược định hướng sản xuất kinh doanh của các doanh nghiệp trong lĩnh vực công nghiệp nội dung thông tin,... sẽ ngày càng được khẳng định.

Định luật Zipf đề cập tới *thứ hạng và tần số* xuất hiện các từ trong một văn

bản/tài liệu. Nội dung định luật:

Trong một văn bản có độ dài đủ cần, nếu như liệt kê các từ xuất hiện trong đoạn văn bản đó theo trật tự giảm dần của tần số xuất hiện của từ đó thì tích giữa *thứ hạng* của từ đó với *tần số* xuất hiện của từ đó là một hằng số. Công thức tổng quát được G.K. Zipf xây dựng:

$$r \times f = k$$

Trong đó:

- *r* là thứ hạng của từ đang khảo sát,
- *f* là tần số xuất hiện của từ đó trong đoạn văn bản,
- *k* là một hằng số.

Định luật đã được Zipf tìm ra thông qua việc phân tích văn bản tiểu thuyết *Ulysses* của James Joyce. Các kết quả cụ thể mà G.K. Zipf đã khảo sát thống kê như sau:

- Từ có tần số hạng thứ 10 xuất hiện trong văn bản 2.653 lần.
- Từ có tần số hạng thứ 100 xuất hiện trong văn bản 265 lần.
- Từ có tần số hạng thứ 200 xuất hiện trong văn bản 133 lần...

Từ đó ông đi tới kết luận là tích giữa thứ hạng của từ và tần số xuất hiện của nó là một hằng số, ước khoảng 26.500.

Các nghiên cứu sau này để thiết lập cơ chế xử lý tự động hóa thông tin văn bản đều ít nhiều dựa trên các kết quả hay phương pháp nghiên cứu mà G.K. Zipf đã xây dựng và sử dụng. Đặc biệt là các hình thức xử lý như dịch tự động; tìm tin trên các văn bản tự do; nghiên cứu, thiết kế các máy tìm tin trên Internet,....

3. Các khía cạnh nghiên cứu và lĩnh vực ứng dụng (ThS. Cao Minh Kiểm)

Trắc lượng thư mục áp dụng phương pháp định lượng trong nghiên cứu tài liệu khoa học. Nó dựa vào các dữ liệu có thể đo đếm được trên tài liệu hoặc nguồn tin như:

- Tác giả: số lượng tác giả trong một lĩnh vực, chủ đề, tạp chí, theo cơ quan, quốc gia, v.v..
- Xuất xứ: tỷ lệ xuất bản (chia theo nước, chủ đề,...);
- Cơ quan, quốc gia, ngôn ngữ;
- Nguồn tin: tạp chí, mức độ gia tăng, sự phân tán;
- Tạp chí, nhà xuất bản, sáng chế, ...;
- Nội dung;
- Văn bản, các phần của văn bản, chủ đề,..
- Trích dẫn: được sử dụng nhiều nhất (trích dẫn một tài liệu, được trích trong một tài liệu, đồng trích dẫn (co-citation));
- Những thuộc tính khác có thể định lượng được.

Trắc lượng thư mục có thể được ứng dụng trong các lĩnh vực như [4]:

- Lịch sử khoa học;
- Xã hội học khoa học;
- Chính sách khoa học;
- Phân bố nguồn lực cho khoa học;
- Thông tin, thư viện: chính sách bổ sung, chính sách thanh lọc tài liệu;
- Tổ chức thông tin;
- Quản trị thông tin;
- Lĩnh vực khác.