PHƯƠNG PHÁP ĐO LƯỜNG TRONG NGHIÊN CỨU THƯ MỤC VÀ WEB

ThS. Trần Mạnh Tuấn
Viện Thông tin Khoa học Xã hội

Giới thiệu những vấn đề liên quan đến trích dẫn, về đo lường web (trắc lượng mạng) và Alexa- công cụ đánh giá trang web. Đưa ra một số nhận xét về việc sử dụng các thuật ngữ tiếng Việt thay cho từ tiếng Anh bibliometrics.

1. Một số khái niệm về trích dẫn

Thống kê thư mục sử dụng các phương pháp khác nhau của phân tích trích dẫn nhằm thiết lập các mối quan hệ giữa các tác giả, các công trình của họ và việc sử dụng các công trình khoa học. Dưới đây xin giới thiệu một số khái niệm về trích dẫn.

Trước hết, *trích dẫn* là việc một tác giả sử dụng, viện dẫn kết quả nghiên cứu đã có ở tài liệu khác trong công trình của mình. Danh mục các tài liệu tham khảo của một công trình khoa học, các chú thích cuối trang (kiểu footnote) liên quan tới tài liệu hay ý kiến, quan điểm của người khác, sự viện dẫn trực tiếp quan điểm nào đó trong văn bản các tài liệu,... là những hình thức trích dẫn rất phổ biến từ trước tới nay.

Phân tích trích dẫn. Khi một tác giả trích dẫn một tác giả khác, tức là mối quan hệ đã được thiết lập. Phân tích trích dẫn sử dụng các trích dẫn trong các công trình khoa học để thiết lập các mối liên kết. Có rất nhiều các mối liên kết khác nhau có thể được xác định như mối liên kết giữa các tác giả, mối liên kết giữa các công trình khoa học, giữa các tạp chí, giữa các lĩnh vực khoa học,... Cả hai loại liên kết trích dẫn bao gồm trích dẫn tới (tài liệu trích dẫn đến tài liệu khác - citing document) hoặc được trích dẫn (tài liệu được tài liệu khác trích dẫn - cited document) đều được nghiên cứu, khảo sát. Phân tích trích dẫn được sử dụng để xác định vai trò, vị trí của một tác giả/tác phẩm/công trình khoa học trong một lĩnh vực cụ thể. Điều này được thể hiện thông qua số lần mà các tác giả khác trong lĩnh vực đó đã trích dẫn đến các công trình của tác giả này.

Liên kết đồng trích dẫn (co-citation coupling) là kiểu liên kết được xác định thông qua việc các tài liệu khác nhau cùng được tài liệu/các tài liệu khác trích dẫn tới - qua đó giữa các tài liệu này đã tồn tại sự tương tự về chủ đề nội dung. Nếu tài liệu A và tài liệu B cùng được một tài liệu C trích dẫn tới thì có thể nói rằng, các tài liệu A và B cùng liên quan đến một chủ đề nào đó, ngay cho dù chúng có thể không trực tiếp trích dẫn với nhau. Nếu các tài liệu A và B kể trên càng được *nhiều* tài liệu khác trích dẫn đến thì mối quan hệ giữa chúng lại càng chặt chẽ hơn. *Số lượng* các tài liệu cùng trích dẫn đến chúng càng lớn *(vấn đề định lượng)* thì mối quan hệ giữa nội dung của chúng lại càng chặt chẽ *(vấn đề định tính)*.

Liên kết thư mục (bibliographic coupling) là kiểu liên kết được xác định thông qua việc các tài liệu khác nhau cùng trích dẫn đến tài liệu/các tài liệu khác - qua đó giữa các tài liệu này cũng tồn tại sự tương tự về chủ đề

nội dung. Ví dụ có 2 tài liệu A và B cùng trích dẫn đến tài liệu C, khi đó có thể nói chúng được liên kết với nhau (hiểu theo ý là giữa chúng có điểm chung nhất định về chủ đề nội dung), thậm chí ngay cả khi chúng không trực tiếp trích dẫn đến nhau. Nếu *số lượng* tài liệu mà chúng cùng trích dẫn đến càng nhiều *(vấn đề định lượng)* thì sự chung nhau về nội dung giữa chúng - mối quan hệ của chúng càng chặt chẽ *(vấn đề định tính)*.

Như đã biết, chỉ dẫn trích dẫn khoa học (Scientific Citation Index - SCI) là một loại sản phẩm thông tin dạng thư mục đặc biệt. Đây là hệ thống tra cứu chỉ dẫn đáp ứng nhu cầu thông tin thư mục về tài liệu. Khác với các loại thư mục khác, SCI bao gồm một hệ thống gồm các loại chỉ dẫn khác nhau để tạo nên một hệ thống bảng tra cứu chỉ dẫn thống nhất. Ở mức tối giản, SCI gồm 3 loại chỉ dẫn: Chỉ dẫn trích dẫn (Citation Index), Chỉ dẫn nguồn (Source Index) và Bảng tra chủ đề hoán vị (Permuterm Subject Index).

Đối tượng được miêu tả trong Chỉ dẫn trích dẫn là các tài liệu được trích dẫn (cited documents), và biểu ghi ứng với mỗi tài liệu này, có liệt kê các tài liệu đã trích dẫn đến.

Đối tượng được miêu tả trong Chỉ dẫn nguồn là các tài liệu trích dẫn (citing documents), và biểu ghi ứng với mỗi tài liệu này, có liệt kê các tài liệu đã được trích dẫn đến.

Đối tượng miêu tả trong Bảng tra chủ đề hoán vị là danh mục các chủ đề mà toàn bộ các tài liệu phản ánh, và tại mỗi chủ đề có liệt kê các tài liệu trích dẫn (nhấn mạnh lại: chúng thuộc loại citing documents đã nêu trên).

Tính chất cơ bản và sự quý giá của SCI chính là ở chỗ việc hệ thống hoá các tài liệu theo dấu hiệu nội dung được phản ánh và *xuất phát từ quan điểm của nhà khoa học* với tư cách là người dùng tin - người tạo ra các thông tin khoa học. Đó là điều khác biệt căn bản với các sản phẩm thông tin khác, khi mà sự phân nhóm tài liệu xuất phát từ quan điểm và sự hiểu biết của cán bộ thông tin thư viện chuyên nghiệp. Nếu như hiện nay, một trong những nguyên tắc cơ bản trong quá trình tạo lập sản phẩm và dịch vụ của các cơ quan thông tin thư viện là định hướng người dùng, thì rõ ràng cách SCI tạo nên là rất đáng được quan tâm [9].

SCI có chức năng kiểm soát tài liệu là bài trích trên một danh sách tạp chí nguồn và các tài liệu đã được các tài liệu này trích dẫn/tham khảo. Đây là một loại sản phẩm khá đặc biệt, được biên soạn và xuất bản tại một số nước có trình độ khoa học phát triển và nguồn tài liệu phong phú như Mỹ, Anh, Đức, Pháp, Nga, Trung Quốc, Ấn Độ,... Hiện tại, khi truy cập đến đa phần các nguồn thông tin toàn văn và trực tuyến của các cơ quan xuất bản và kinh doanh tài liệu khoa học trên thế giới dễ dàng có thể khai thác, sử dụng các hệ thống chỉ dẫn trích dẫn giúp người dùng tìm kiếm được các công trình nghiên cứu có giá trị thông qua số lượng các công trình đã trích dẫn đến tài liệu đó.

Gần đây, người ta thường nhắc tới danh mục các tạp chí khoa học có uy tín trên thế giới (Danh mục tạp chí khoa học Philadelphia). Số lượng các công trình khoa học của một cá nhân/tổ chức khoa học được công bố trên danh sách đó được coi là chỉ số quan trọng và tin cậy để xác định vai trò và uy tín khoa học của cá nhân/cộng đồng đó. Dù còn một số ý kiến khác biệt nhau, song chỉ số này hiện vẫn được thừa nhận là một trong các công cụ quan trọng để so sánh trình độ nghiên cứu khoa học của các quốc gia, cộng đồng, cá nhân. Danh mục tạp chí khoa học Philadelphia được xây dựng trên cơ sở lựa chọn các tạp chí khoa học có số lượng

các trích dẫn khoa học giảm dần. Bởi vậy, các CSDL về các chỉ dẫn trích dẫn khoa học sẽ trở thành công cụ kiểm định và lựa chọn tạp chí nào sẽ lọt vào danh mục này và ngược lại.

2. Phương pháp đo lường web - sự tiếp tục của phương pháp đo lường thư mục

Gần đây, một lĩnh vực mới xuất hiện và phát triển mạnh có liên quan đến thống kê thư mục được gọi là **webometrics** hoặc **cybermetrics**. Đo lường web có thể được định nghĩa như là việc sử dụng các kỹ thuật/phương pháp của thống kê thư mục nhằm nghiên cứu mối quan hệ giữa các trang web. Các kỹ thuật hay phương pháp này có nhiệm vụ xác định số lần mà mỗi trang web kết nối đến các trang web khác.

Trong các toán tử được sử dụng trong máy tìm tin đơn Altavista, có một toán tử khá đặc biệt với cú pháp thực hiện:

Link: <URL>

Trong đó, URL là địa chỉ của một trang web cụ thể.

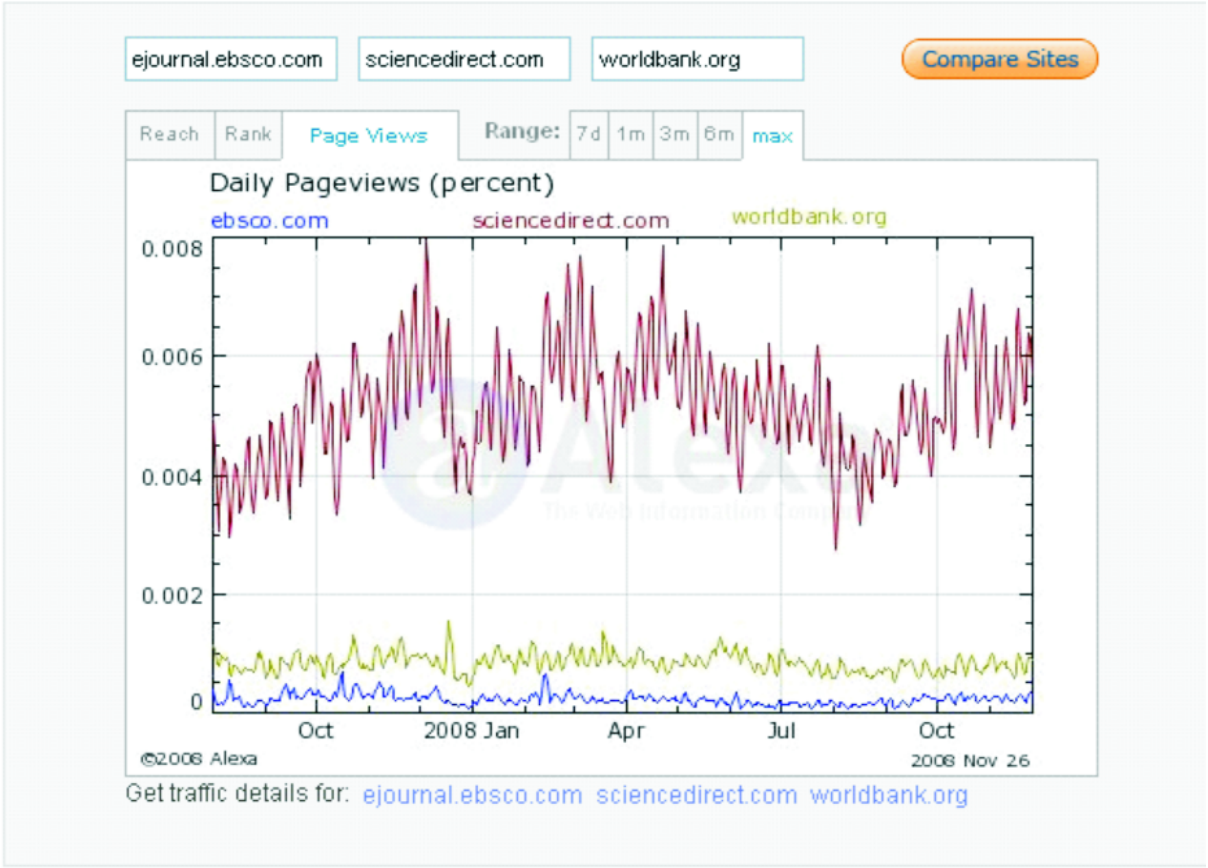
Khi đó, kết quả nhận được sẽ là danh mục các trang web có xây dựng các liên kết đến trang web có địa chỉ URL trong lệnh tìm trên. Tại thời điểm ngày 23/7/2008, thực hiện lệnh tìm *link: vista.gov.vn* trên altavista, nhận được kết quả là có 5.450 trang web tại Mỹ có thiết lập các liên kết đến http://www.vista.gov.vn

Rõ ràng, khi một trang web có càng nhiều trang web khác xây dựng các đường liên kết đến (trích dẫn đến), thì vai trò, vị trí của trang web đó càng được khẳng định trong cộng đồng mạng. Các kết quả thống kê cho phép kết luận điều đó, và đương nhiên các nhà quản trị các trang web sẽ phải luôn nỗ lực và đổi mới để sao cho kết quả thực hiện lệnh tìm trên đối với trang web của mình ngày càng lớn.

Khi nói đến đo lường web, trước hết cần quan tâm đến sự xuất hiện của một công cụ đang được phổ biến hiện nay là Alexa với địa chỉ: http://www.alexa.com. Đây là công cụ đánh giá trang web dựa trên hai chỉ số chính là số trang web được người dùng xem (page view) và số lượng người truy cập trên trang web đó (page reach). Alexa chỉ tính toán dựa trên các máy tính mà trình duyệt web có tích hợp thanh công cụ Alexa Toolbar. Hiện có khoảng khoảng 1% số người dùng Internet sử dụng Alexa Toolbar.

Các chỉ số mà Alexa thống kê đối với mỗi trang web đã trở thành các tiêu chí thể hiện vai trò của nó đối với người dùng tin nói chung. Ngoài Alexa, hiện có khá nhiều công cụ khác có thể được sử dụng để đánh giá và xếp hạng trang web như Compete, ComScore, Hitwise, Nielsen// NetRatings, Netcraft, Ranking.com, Quantcast... Ở nước ta, Alexa là công cụ được sử dụng phổ biến nhất để thực hiện *đo lường và thống kê* về việc khai thác, truy cập các trang web (dựa trên các số liệu xếp hạng trang web - trafic rankings).

Trang chủ của Alexa cung cấp các thông tin dưới dạng đồ thị so sánh các trang web. Để có các thông tin so sánh, cần điền vào hộp hội thoại địa chỉ các trang web cần được so sánh. Trong hình là việc so sánh theo số lượng lượt xem các trang web (page views) giữa 3 trang web quen thuộc có địa chỉ: ejournal.ebsco.com, sciencedirect.com và worldbank.org, theo các số liệu thống kê trong khoảng thời gian từ tháng 8/2007 trở đi (đến 26/11/2008). Ngoài ra, các tiêu chí so sánh có thể được lựa chọn là về việc truy cập (reach), thứ hạng (rank) và số lượt xem trang (page views) và theo các khoảng thời gian hồi cố khác nhau tính tới thời điểm hiện tại: 1 tuần, 1 tháng, 3 tháng, 6 tháng và trên 1 năm.



So sánh thứ hạng một số trang web trong khoảng 1 năm (từ tháng 8/2007 đến 26/11/08)

Các trang web được Alexa xếp hạng theo 3 cách phân nhóm:

- Xếp hạng chung trên toàn cầu (global): cho phép tìm các trang web theo thứ hạng giảm dần. Ở đây không phân biệt xuất xứ và lĩnh vực hoạt động mà trang web đó trực thuộc.

- Xếp hạng theo mỗi quốc gia (by country): Hiện tại đã thống kê đối với các trang web của 134 quốc gia.

- Xếp hạng theo ngôn ngữ (by language). Hiện tại, Alexa đã thống kê các trang web được thể hiện thông qua 21 ngôn ngữ, bao gồm: Ngôn ngữ Ảrập, Tiếng Trung Quốc giản lược, Tiếng Trung Quốc truyền thống, Séc, Đan mạch, Đức, Anh, Phần Lan, Pháp, Hà Lan, Hy Lạp, Hebrew, Ý, Nhật Bản, Hàn Quốc, Nauy, Bồ Đào Nha, Nga, Tây Ban Nha, Thuy Điển và Thổ Nhĩ Kỳ.

- Xếp hạng theo các lĩnh vực hoạt động (by category). Hiện tại Alexa hệ thống hoá các trang web theo 16 lĩnh vực hoạt động bao gồm: Nghệ thuật (254.761 trang), Thương mại (234.476), Máy tính (115.087), Các trò chơi (56.461), Sức khoẻ (60.987), Gia đình và nhà cửa (28.487), Trẻ em và lứa tuổi vị thành niên (44.690), Thời sự, tin tức (8.732), Giải trí (106.152), Tra cứu-chỉ dẫn tham khảo (57.625), Các trang web về các vùng, khu vực trên thế giới (1.084.196), Các ngành khoa học (105.390), Mua sắm

(96.101), Xã hội (238.571), Thể thao (100.610), Thế giới nói chung (1.722.130).

Là một công cụ mạnh và sắc bén, tuy nhiên, cách thức xác lập các số liệu thống kê mà Alexa thực hiện rất dễ bị lợi dụng do rất khó xác định nổi đâu là số liệu thực, đâu là các số liệu ngụy tạo. Có thể dẫn ra một hình ảnh về sự lợi dụng để khuếch đại số lượt truy cập đến một trang web là số lượng người hiện đang sinh sống của một thành phố và số cư dân của thành phố đó. Một số nhà quản trị các trang web vô danh tiểu tốt đã tạo liên kết trang web này đến các trang web danh tiếng khác và xác định số liệu thống kê về lượt truy cập đến trang web của mình bằng tổng số lượt truy cập đến tất cả các trang web đã được *nó* kết nối tới (!).

3. Thử đi tìm một tên gọi

Việt Nam là một nước mà trình độ và truyền thống về khoa học, trong đó có khoa học thư viện và thông tin, chưa đạt mức tiên tiến trên thế giới. Những cán bộ nghiên cứu trong lĩnh vực này chủ yếu hoạt động trên cơ sở thu nhận và ứng dụng các kết quả, thành tựu mà các học giả ở nước ngoài đã nghiên cứu và khám phá. Các biểu hiện và tác động của trình độ phát triển khoa học ở nước ta đa dạng và phong phú, trong đó trước hết là:

- Hệ thống các thuật ngữ, khái niệm khoa học tiếng Việt chưa phát triển kịp và phản ánh được đầy đủ các thành tựu khoa học trong các lĩnh vực này;

- Hệ thống thuật ngữ khoa học chưa được sử dụng một cách thống nhất, chưa đủ hàm súc và chưa mang tính hệ thống;

- Trong nhiều trường hợp, các nhà nghiên cứu buộc phải trực tiếp sử dụng các thuật ngữ nước ngoài như bibliometrics, webometrics, cybermetrics, website, WWW[3],...

Về nguồn gốc sâu xa, vĩ tố- *metrics* được vay mượn từ một khái niệm cơ bản trong toán học. Trong các khoa học thư viện và thông tin, khi ghép vĩ tố này với các thuật ngữ thư mục (*biblio*), tài liệu trên mạng (*web*) là để:

- Ám chỉ một phương pháp nghiên cứu sử dụng các số liệu thống kê, các thông tin mang tính định lượng.

- Nhằm mục đích xác định giá trị hoặc rộng hơn là quan hệ của một *tài liệu* trong tập hợp (không gian) các tài liệu.

- Khảo sát và dự báo xu hướng (bao gồm cả *cường độ* và *động thái*) phát triển nguồn tin.

Xây dựng nền tảng khoa học - cơ sở phương pháp luận đối với việc xử lý thông tin nói chung, đặc biệt là xử lý thông tin tự động hoá, xử lý thông tin tồn tại dưới dạng số.

Với quan điểm và cách tiếp cận như trên, theo thiển ý của chúng tôi, có thể lựa chọn để sử dụng một trong số các cách sau đây: Bibliometrics/Webometrics/Cybermetrics; Đo lường thư mục (web); Trắc lượng thư mục (web); Phương pháp đo lường thư mục (web); Phương pháp trắc lượng thư mục (web); Phương pháp đo lường trong nghiên cứu thư mục (web); Phương pháp trắc lượng trong nghiên cứu thư mục (web). Bảng dưới đây đưa ra một số so sánh trong việc sử dụng các thuật ngữ này.

[3] Đầu tiên, nhan đề của bài viết này được dự định đặt là: Webometrics – sự tiếp tục của bibliometrics. Điều đó cho thấy rõ những khó khăn cho việc lựa chọn một thuật ngữ tiếng Việt tương đương với thuật ngữ tiếng Anh được sử dụng rất phổ biến ở nước ngoài.

TT	Thuật ngữ được chọn	Sơ bộ đánh giá	Ghi chú
1	Sử dụng các thuật ngữ gốc hoặc biến thể (phiên âm sang tiếng Việt) *Bibliometrics/ Webometrics (Cybermetrics)*	Ưu điểm: Tạo nên sự thống nhất trong việc sử dụng; hàm súc; chính xác; phù hợp với xu thế hội nhập,.... Hạn chế: Vay mượn thuật ngữ nước ngoài; không kích thích sự phát triển khoa học; việc phổ cập bị hạn chế; dễ tạo tâm lí bài xích đối với người dùng,....	
2a	*Phương pháp đo lường thư mục*	Ưu điểm: Đã phản ánh được bản chất chính của phương pháp nghiên cứu; dễ phổ biến tại Việt Nam; góp phần phát triển thuật ngữ khoa học tiếng Việt,.... Hạn chế: Về từ nguyên, chưa phản ánh đầy đủ các nội dung của phương pháp nghiên cứu; chưa nhanh chóng tạo được sự thống nhất trong sử dụng	Thay thế Bibliometrics
2b	*Phương pháp trắc lượng thư mục*	Tương tự như trên, song đây là thuật ngữ Hán Việt	Thay thế: Bibliometrics
3a	*Phương pháp đo lường trong nghiên cứu thư mục*	Ưu điểm: Đã phản ánh được bản chất chính của phương pháp nghiên cứu; dễ phổ biến tại Việt Nam; góp phần phát triển thuật ngữ khoa học tiếng Việt; so với thuật ngữ trên, đã phản ánh được đầy đủ hơn về nội dung,.... Hạn chế: Quá cồng kềnh nên rất khó sử dụng thống nhất.	Thay thế: Bibliometrics
3b	*Phương pháp trắc lượng trong nghiên cứu thư mục*	Tương tự tình huống 3a, song đây là thuật ngữ Hán Việt	Thay thế: Bibliometrics
4a	*Phương pháp đo lường web*	Tương tự trường hợp thứ 2a	Thay thế webometrics (cybermetrics)
4b	*Phương pháp trắc lượng web*	Tương tự tình huống 2a	Thay thế webometrics (cybermetrics)
5a	*Phương pháp đo lường trong nghiên cứu web*	Tương tự trường hợp thứ 3a	Thay thế webometrics (cybermetrics)
5b	*Phương pháp trắc lượng trong nghiên cứu web*	Tương tự trường hợp thứ 3b	Thay thế webometrics (cybermetrics)

4. Kết luận

Cùng với xu thế phát triển các ứng dụng công nghệ thông tin và truyền thông (ICT) vào công tác thông tin-thư viện, các hướng nghiên cứu cơ bản trong lĩnh vực này cũng được quan tâm, bởi vấn đề của thông tin- thư viện không chỉ là các khía cạnh liên quan tới công nghệ. Một trong số các vấn đề lớn và phức tạp ở đây chính là phải hiểu rõ và đầy đủ quy luật phát triển các loại nguồn tin, mối quan hệ của chúng với quá trình phát triển khoa học; phải hiểu rằng quá trình nguồn tin được hình thành, được khai thác sử dụng là một quá trình diễn ra theo đường xoắn ốc với tần số ngày càng cao, biên độ ngày càng lớn, không còn tuân theo quy luật phát triển *hàm số mũ* như trước đây người ta đã dự báo, bởi các tác nhân chính gây nên hiện tượng bùng nổ thông tin đã có những khác biệt căn bản so với trước đây.

Với cách đặt vấn đề như vậy, chúng tôi hy vọng bài viết này thu hút được sự quan tâm của các đồng nghiệp, trước hết là những người làm công tác nghiên cứu, giảng dạy tại các viện nghiên cứu, trường đại học. Có thể thấy đây là vấn đề hết sức quen thuộc và là bản chất của các khoa học thư viện và thông tin: quy luật và tính chất phát triển nguồn thông tin và việc sử dụng thông tin.

Tài liệu tham khảo

1. Bibliometrics, http://en.wikipedia.org/wiki/Thống kê thư mục.

2. Bibliometrics, http://www.ischool.utexas.edu/~palmquis/courses/biblio.html

3. Dictionary of bibliometrics, http://books.google.com.vn/books?hl=en&id=XBg1SNzNTD0C&dq=thống kê thư mục&printsec=frontcover&source=web&ots=FpmoNntyV_&sig=qUQiIfAt05kJbH4REDhW-XHryvs&sa=X&oi=book_result&resnum=6&ct=result

4. Elliott Amy J., Self- and Cross-Citations, in the Journal of Applied Behavior Analysis and the Journal of the Experimental Analysis of Behavior: 1993–2003

5. Garfield E., The Concept Citation Indexing: Unique and Innovative Tool for Navigating the Research Literature, http://www.thomsonreuters.com/business_units/scientific/free/essays/history/

6. Harboe-Ree C., Bibliometrics information kit, http://www.caul.edu.au/stats/caul20052bibliometrics.doc

7. Glanzel W., Bibliometrics as a research field: A course on theory and application on thống kê thư mục indicators, 2003 http://www.norslis.net/2004/Bib_Module_KUL.pdf

8. Library Science: Information Science:, Bibliometrics http://groups.yahoo.com/group/Net-Gold/message/13439

9. Nzel W.G., Bibliometrics as research field, http://www.norslis.net/2004/Bib_Module_KUL.pdf

10. Rios D.R,. The bibliometrics: penetration level in the university teaching of library science and its application in the librarian field in the countries of Mercosur, http://www.ifla.org/IV/ifla66/papers/162-127e.htm

11. Sara von Ungern-Sternberg, Applications in teaching Bibliometrics, http://www.ifla.org/IV/ifla61/61-ungs.htm

12. Webometrics At the: http://en.wikipedia.org/wiki/Thống kê web

13. Thelwall M, Bibliometrics to webometrics, http://jis.sagepub.com/cgi/content/abstract/34/4/605

14. Webometrics – Search engine, web crawler and web link analysis information, http://webometrics.blogspot.com

15. What's the IMPACT of Your Research? Bibliometric Applications in Evaluating Research: Concepts, Tools and Analyses, https://wis.ntu.edu.sg/pls/webexe/REGISTER_NTU.REGISTER?EVENT_ID=OA07032816584500.

16. Nghiên cứu, xây dựng hệ thống sản phẩm, dịch vụ thông tin tại Viện Khoa học xã hội Việt Nam giai đoạn hiện nay: Đề tài nghiên cứu khoa học cấp Bộ./ Chủ nhiệm: Trần Mạnh Tuấn, Viện Khoa học xã hội Việt Nam, H., 2008, 170 tr.