CƠ SỞ DỮ LIỆU TRẮC LƯỢNG THƯ MỤC[1]

TS Nguyễn Huy Chương
Khoa TT-TV, Trường Đại học KHXH&NV, ĐHQG Hà Nội
PGS TS Đỗ Trung Tuấn
Khoa Toán Cơ Tin, Trường Đại học KHTN, ĐHQG Hà Nội

Tóm tắt: *Trắc lượng thư mục có ý nghĩa đánh giá công trình nghiên cứu khoa học đối với người nghiên cứu, đồng thời, thể hiện năng lực của tổ chức nghiên cứu khoa học và tổ chức quản lý khoa học. Để triển khai hoạt động này, cần có cơ sở dữ liệu, cho phép cung cấp thông tin để thực hiện đo lường/đánh giá. Bài viết phân tích và đề xuất, thiết kế cơ sở dữ liệu với SQL server, nhằm hỗ trợ việc tổ chức, lưu trữ và xử lý thông tin liên quan đến trắc lượng thư mục.*

Từ khóa: Cơ sở dữ liệu; trắc lượng thư mục; nghiên cứu khoa học; SQL; Đại học Quốc gia Hà Nội.

Bibliometric databases

Abstract: *Bibliometrics is used to evaluate the work of a researcher as well as the capacity of a research institution and a research management organization. In order to conduct bibliometric analysis, it's necessary to have databases to provide information for monitoring and evaluation. The article analyzes the current status of bibliometric databases and recommends to design bibliometric database with SQL server in order to organize, store and analyze bibliometric-related information.*

Keywords: Databases; bibliometrics; scientific research; SQL; Vietnam National University Hanoi.

1. Đặt vấn đề

Đánh giá chất lượng sản phẩm thông tin cần có dạng cơ sở dữ liệu chuyên dụng. Thực tế cho thấy, ở Việt Nam hiện nay thiếu cơ sở dữ liệu trắc lượng thư mục phù hợp. Để khắc phục vấn đề này, cần phải nhờ đến một giải pháp đã được thực hiện từ nhiều năm trước, bao gồm: (i) tải dữ liệu; (ii) làm sạch nó; và (iii) lưu trữ nó vào một cơ sở dữ liệu thích hợp cho các nhiệm vụ trắc lượng thư mục. Đối với các đơn vị nghiên cứu, vấn đề là làm thế nào một cơ sở dữ liệu như vậy được xây dựng để đáp ứng tốt nhất nhu cầu trắc lượng thư mục [1].

Để việc đánh giá tiện cho người dùng không chuyên công nghệ thông tin, giao diện người dùng cần thân thiện, phù hợp. Nhiều tiêu chí đặt ra đối với giao diện người-máy; nhưng với hệ thống trắc lượng thư mục, cần có các tiêu chí phù hợp với công tác TT-TV và hệ thống cần có phần tương tác người dùng theo cách trực quan [2]. Vì vậy, mục đích của bài viết này nhằm mô tả cấu trúc của một cơ sở dữ liệu quan hệ, thích hợp cho hầu hết các phân tích, thiết kế và tính toán các chỉ số trắc

1 Bài viết được thực hiện khi tiến hành đề tài nghiên cứu được tài trợ bởi Đại học Quốc gia Hà Nội

lượng thư mục.

Gần đây, nhu cầu về cơ sở dữ liệu chuyên dùng cho các mục đích trắc lượng thư mục đã được khẳng định và xuất hiện hướng mới trong mô hình dữ liệu. Người ta có thể sử dụng tiếp cận quan hệ - đối tượng. Tuy nhiên, điều này không có nghĩa là cơ sở dữ liệu quan hệ thuần túy là lỗi thời, mà việc sử dụng công nghệ hướng đối tượng nhằm thích hợp hơn với hoạt động trắc lượng thư mục.

Bài viết này nhằm mô tả cấu trúc của một cơ sở dữ liệu quan hệ, thích hợp cho hầu hết các phân tích, thiết kế và tính toán các chỉ số trắc lượng thư mục. Trong khi vẫn chưa có một cơ sở dữ liệu quan hệ mẫu, bài viết sẽ phân tích, thiết kế một cơ sở dữ liệu quan hệ phù hợp với công tác trắc lượng thư mục tại Đại học Quốc gia Hà Nội

2. Phân tích cơ sở dữ liệu trắc lượng thư mục

Phân tích sử dụng chỉ số trắc lượng thư mục có thể được phân thành: (i) *trắc lượng thư mục mô tả*; (ii) *trắc lượng thư mục đánh giá*. Trong khi trắc lượng thư mục mô tả theo tiếp cận từ trên xuống, cố gắng để có được những bức tranh lớn, chẳng hạn kết quả nghiên cứu của một quốc gia trong các lĩnh vực khác nhau, tỷ lệ của các lĩnh vực khác nhau và thay đổi theo thời gian, thì trắc lượng thư mục là một công cụ để đánh giá hoạt động nghiên cứu của các đơn vị nhỏ hơn như nhóm nghiên cứu hoặc thậm chí các cá nhân và sử dụng một phương pháp tiếp cận từ dưới lên, thu thập tất cả các ấn phẩm (có liên quan) của các đơn vị tương ứng. Rõ ràng, trắc lượng thư mục đánh giá đặt ra yêu cầu cao hơn về chất lượng dữ liệu.

Tính toán các chỉ số trắc lượng thư mục là đếm số ấn phẩm và trích dẫn. Liên quan đến vấn đề này có một số ý kiến sau:

• Công việc này đề cập con số định lượng, nhưng vấn đề quan trọng là đảm bảo chất lượng dữ liệu. Chất lượng dữ liệu được xác định qua các đặc trưng, tức các từ khóa mà người ta lựa chọn để thống kê [3];

• Một số đặc trưng liên quan đến con người, như tiểu sử cá nhân, cũng được xem xét, khảo cứu để tra cứu, đánh giá công trình [4];

• Tác động của một công trình này đến công trình khác là tác động trực tiếp. Tuy nhiên, công trình thứ hai lại có ảnh hưởng đến công trình thứ ba, thứ tư… Do vậy, việc truy vết tác động của một công trình cũng cần được thể hiện trong cơ sở dữ liệu. Kinh nghiệm cho thấy cần xác định đường đi của một công trình trong mạng lưới các công trình [7];

• Trong cơ sở dữ liệu về trắc lượng thư mục, các đối tượng chính được coi như đặc trưng xác định các đặc trưng khác, chẳng hạn thuộc tính khóa trong cơ sở dữ liệu. Nên xác định tên đối tượng số hóa trong hệ thống đánh giá có uy tín, như trên Web về khoa học và Scopus [5].

2.1. Trắc lượng thư mục

Trong bài này, chúng tôi sử dụng một số tiêu chí về chỉ số trắc lượng thư mục trên cơ sở tham khảo các tiêu chí của Nicolai Mallig [6] với các ký pháp. Việc sử dụng lại các ký pháp nhằm thuận tiện cho việc đối chiếu, so sánh.

• *P*. Số lượng ấn phẩm;

• *C*. Số trích dẫn nhận được;

• *CPP*[2]. (Số trung bình) các trích dẫn đối với mỗi ấn phẩm;

• *CPPex*[3]. (Số trung bình) trích dẫn đối với mỗi xuất bản; không tính tự trích dẫn;

2 CPP : Citation per publication
3 CPPex : CPP, self citation excluded

• *% Pnc*. Tỷ lệ phần trăm của các bài báo không được trích dẫn (trong khoảng thời gian xem xét);

• *JCS*[4]. Tỷ số trích dẫn tạp chí (số trung bình của các trích dẫn trên mỗi xuất bản, theo loại bài báo và tạp chí);

• *FCS*[5]. Tỷ số trích dẫn lĩnh vực (số trung bình của các trích dẫn trên mỗi xuất bản, theo loại tạp chí và lĩnh vực hẹp);

• *JCSm*. Tỷ lệ trích dẫn trung bình của gói tạp chí (đánh trọng số theo số lượng các ấn phẩm của gói bài báo đang xét);

• *FCSm*. Tỷ lệ trích dẫn trung bình theo lĩnh vực (đánh trọng số theo số lượng các ấn phẩm của gói các bài báo đang xét);

• *% SELFCIT*. Tỷ lệ tự trích dẫn;

• *CPP/JCSm*. Trích dẫn theo xuất bản, so với tỷ lệ trích dẫn của các gói tạp chí;

• *CPP/FCSm*. Trích dẫn theo xuất bản, so với tỷ lệ trích dẫn của các lĩnh vực hẹp;

• *JCSm/FCSm*. Tỷ lệ trích dẫn của các gói tạp chí, so với tỷ lệ trích dẫn các lĩnh vực hẹp.

Nicolai Mallig [6] đề xuất thêm một chỉ số khác, thường được sử dụng như đánh giá hợp tác (quốc tế), đó là chỉ số: CoP. Số cùng xuất bản (cùng với một đơn vị khác).

Đếm ấn phẩm là hoạt động nhằm *tính điểm* cho ấn phẩm. Các điểm được gộp lại, theo các mức độ, chẳng hạn theo tác giả, tổ chức hay quốc gia. Những phương pháp đếm thường được sử dụng gồm:

1. Toàn bộ[6]. Mỗi đơn vị cơ bản có liên quan (tác giả) được thêm 1 điểm;

2. Toàn bộ theo chuẩn[7]. Mỗi tác giả tham gia được điểm 1/n, với n là số các tác giả của bài báo;

3. Trực tiếp[8]. Tác giả đầu tiên nhận được 1 điểm; tác giả khác không được tính điểm;

4. Tất cả[9]. Mỗi đơn vị tham gia, theo mức độ gộp lại, đều được điểm 1;

5. Tất cả bình thường[10]. Mỗi đối tượng có liên quan, theo mức độ gộp lại, tăng thêm 1/n điểm; với n là số lượng đơn vị tham gia, theo mức độ này.

Hai phương pháp hoàn toàn bình thường, tự nhiên, là các loại phương pháp đếm một phần, hay toàn phần. Chẳng hạn bài báo với hai tác giả Việt Nam và một tác giả người Pháp.

• Nếu đếm một phần, tác giả Việt Nam được 2/3 điểm; tác giả Pháp được 1/3 điểm;

• Nếu đếm toàn phần, tức đầy đủ, ½ điểm cho Việt Nam, và ½ điểm cho Pháp.

Về các cơ sở dữ liệu để phân tích trắc lượng thư mục hiện có trên thế giới, người ta thường kể đến: (i) ISI Web of Science (Thomson Reuters); (ii) Scopus (Elsevier); (iii) Google Scholar (Google Inc); và (iv) Các thư mục cục bộ.

2.2. Cơ sở dữ liệu bổ sung, nhằm truy vết các trích dẫn

Truy vết các ấn phẩm là cần thiết [7]. Một số nhà xuất bản hoặc cơ sở dữ liệu cung cấp thông tin theo dõi trích dẫn. Dưới đây là danh sách một số cơ sở dữ liệu như vậy.

1. Thư viện kỹ thuật số ACM: CSDL gồm

4 JCS : Journal Citation Score
5 FCS : Field Citation Score
6 Complete
7 Complete normalized
8 Straight
9 Whole
10 Whole normalized

các bài báo và các hội nghị khoa học máy tính và công nghệ thông tin. Việc tìm kiếm một tác giả hoặc công việc cụ thể là dễ dàng. Đối với mỗi công việc, số lượng trích dẫn và số lượng tải được hiển thị;

2. IEEE Xplore: Gồm các bài báo và kỷ yếu hội nghị về công nghệ và khoa học máy tính;

3. MathSciNet: Gồm các bài báo, kỷ yếu hội nghị, và sách về toán học;

4. Tạp chí ScienceDirect: Gồm các bài báo về y học, khoa học, một số ngành khoa học xã hội, nhân văn,….

3. Đề xuất lược đồ cơ sở dữ liệu quan hệ

Để xác định các thành phần cấu trúc chính của một bài báo và các mối quan hệ của chúng, cần xem xét các khái niệm tổng quan liên quan đến bài báo, bao gồm:

• Bài viết có tiêu đề và tóm tắt.

• Bài báo được viết bởi một hoặc nhiều tác giả (thứ tự xuất hiện có thể là thông tin quan trọng).

• Tác giả liên kết với một tổ chức (hoặc một số) trong đó có địa chỉ.

• Tác giả có thể có địa chỉ e-mail.

• Bài viết được đăng tải trên một tạp chí có tên.

• Bài viết được xuất bản theo vấn đề cụ thể của tạp chí. Đặc trưng của nó là *chất lượng, ngày xuất bản*.

• Bài viết này có ngày tác giả gửi và ngày tạp chí nhận.

• Có một số từ khóa (được cung cấp bởi các tác giả).

• Bài báo có một danh sách các tài liệu tham khảo đến các bài báo khác.

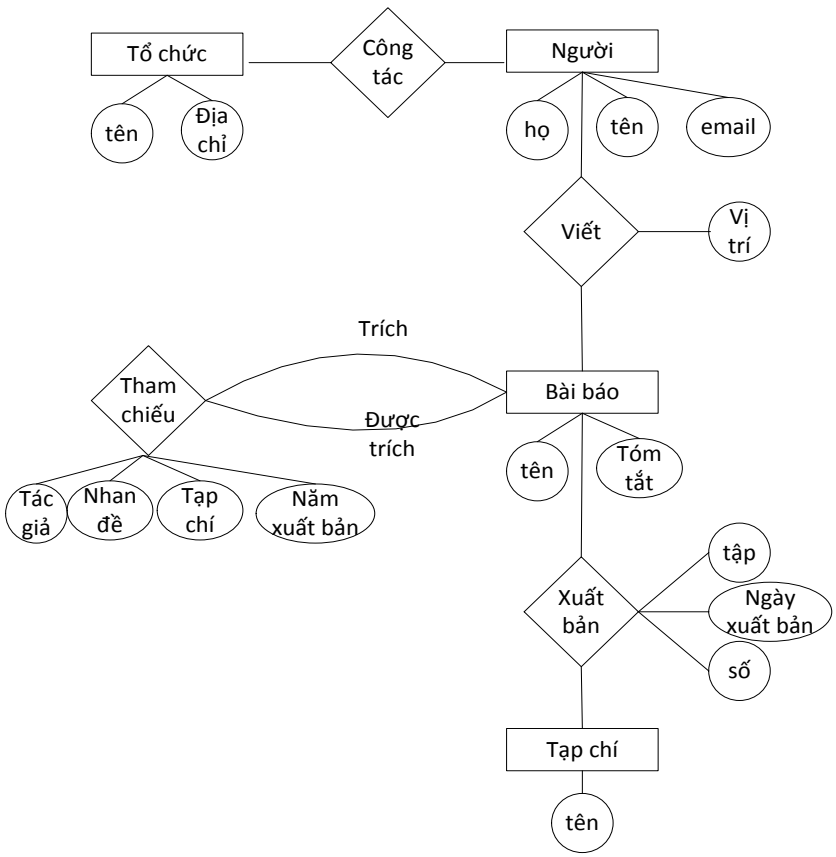
• Mỗi tài liệu tham khảo có chứa thông tin đầy đủ (trong trang đầu của bài báo trích dẫn).

Liên quan đến tổ chức thông tin, có các đối tượng quan trọng như: (i) bài báo; (ii) tạp chí; (iii) người (tác giả); (iv) cơ quan.

Các mối quan hệ được xác định, tức các thực thể liên kết, là:

• tác giả (liên kết người và bài viết);

• xuất bản (liên kết bài báo và tạp chí);



Hình 1. *Sơ đồ ER cơ bản*

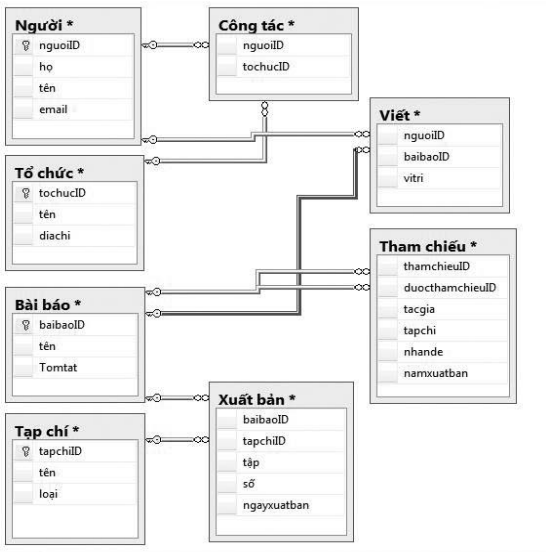
• liên kết (liên kết người và tổ chức);

• tài liệu tham khảo/trích dẫn (liên kết bài viết với bài viết, liên kết trích dẫn với trích dẫn).

Ở đây sử dụng mô hình thực thể- quan hệ (ER) để hình dung các thực thể với các thuộc tính và các mối quan hệ xác định ở trên. Một sơ đồ thực thể- quan hệ là một thể hiện trừu tượng của dữ liệu, thường được sử dụng để mô hình hóa dữ liệu.

Các đối tượng được hiển thị như: (i) hình hộp ứng với thực thể; (ii) thuộc tính ứng với hình bầu dục; (iii) mối quan hệ ứng với hình thoi; (iv) các mũi tên với nhãn.

Các thực thể và các mối quan hệ được xác định được hiển thị trong Hình 1. Thực tế là một bài báo được công bố trên một tạp chí được thể hiện thông qua các mối quan hệ liên kết xuất bản bài báo cho tạp chí. Các tác giả thiết lập một liên kết giữa một bài báo và người đã viết nó, theo quan hệ viết. Một người thuộc về một tổ chức được thể hiện bởi các mối quan hệ công tác. Các mối quan hệ tài liệu tham khảo liên kết các trích dẫn vào bài báo trích dẫn, tức nó được liên kết hai lần, theo quan hệ tham chiếu.



Hình 2. *Sơ đồ của các lược đồ quan hệ cơ bản*

Column Name	Data Type	Allow Nulls
baibaoID	int	
tên	nchar(10)	
Tomtat	nchar(10)	

Hình 3. *Các thuộc tính của quan hệ Bài báo*

Mối quan hệ *tham chiếu* là phức tạp. Nó có ý nghĩa cho việc đánh giá bản thân bài viết. Vì vậy, các thuộc tính của các mối quan hệ *tham chiếu* là thông tin dự phòng đã được chứa trong dữ liệu của các bài viết được trích dẫn.

Để cụ thể hóa các lược đồ quan hệ, các thực thể, các mối quan hệ và các thuộc tính của mô hình ER phải được tương ứng với các thuộc tính trong các bảng quan hệ của mô hình quan hệ. Sự chuyển đổi này khá đơn giản: các thực thể và các mối quan hệ được ánh xạ vào các bảng trong khi các thuộc tính được ánh xạ vào các cột của bảng tương ứng. Mối quan hệ có thể tạo nên một quan hệ mới.

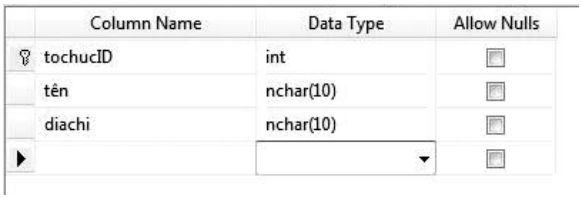
Hệ quản trị cơ sở dữ liệu được đề xuất sử dụng là SQL Server. Đây là hệ quản trị thông dụng, phổ cập đối với mọi cơ quan/trường học/thư viện Việt Nam. Dưới đây là các lược đồ được trình bày theo ngôn ngữ của hệ quản trị này.

Column Name	Data Type	Allow Nulls
nguoiID	int	
họ	nchar(10)	
tên	nchar(10)	
email	nchar(10)	

Hình 4. *Các thuộc tính của quan hệ Người*

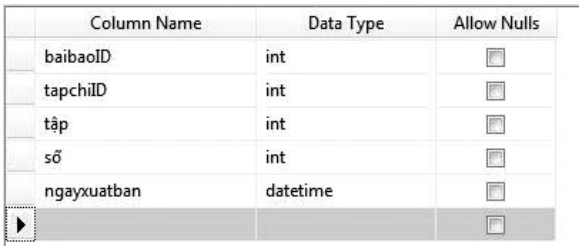
Column Name	Data Type	Allow Nulls
tapchiID	int	
tên	nchar(10)	
loại	nchar(10)	

Hình 5. *Các thuộc tính của quan hệ Tạp chí*



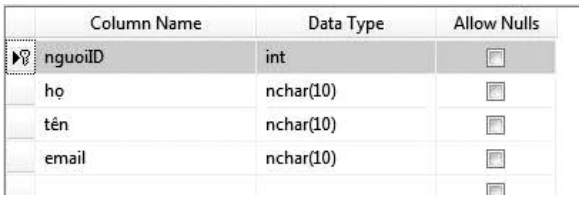
Column Name	Data Type	Allow Nulls
tochucID	int	
tên	nchar(10)	
diachi	nchar(10)	

Hình 6. *Các thuộc tính của quan hệ Tổ chức*



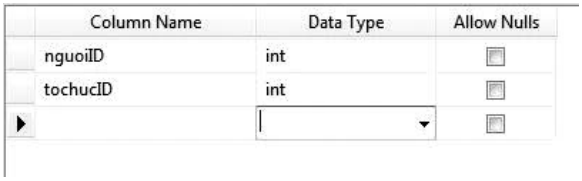
Column Name	Data Type	Allow Nulls
baibaoID	int	
tapchiID	int	
tập	int	
số	int	
ngayxuatban	datetime	

Hình 7. *Các thuộc tính của mối quan hệ Xuất bản*



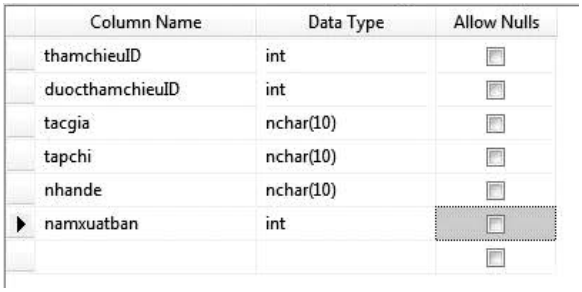
Column Name	Data Type	Allow Nulls
nguoiID	int	
họ	nchar(10)	
tên	nchar(10)	
email	nchar(10)	

Hình 8. *Các thuộc tính của mối quan hệ Viết*



Column Name	Data Type	Allow Nulls
nguoiID	int	
tochucID	int	

Hình 9. *Các thuộc tính của mối quan hệ Công tác*



Column Name	Data Type	Allow Nulls
thamchieuID	int	
duocthamchieuID	int	
tacgia	nchar(10)	
tapchi	nchar(10)	
nhande	nchar(10)	
namxuatban	int	

Hình 10. *Các thuộc tính của mối quan hệ Tham chiếu*

4. Kết luận

Đối với chương trình xử lý dữ liệu, một vấn đề rất quan trọng là dữ liệu cần được thể hiện ở dạng thuận tiện đối với người dùng.

Đề xuất lược đồ cơ sở dữ liệu trong phần trên hoàn toàn có thể đáp ứng được nhu cầu đánh giá, hỗ trợ xếp hạng của trắc lượng thư mục. Lược đồ cơ sở dữ liệu này phù hợp với thực tiễn hoạt động thông tin thư mục tại các trường đại học Việt Nam nên cần được triển khai áp dụng rộng rãi.

TÀI LIỆU THAM KHẢO

1. Andrea Bonaccorsi, Tindaro Cicero (2016). Nondeterministic ranking of university departments, *Journal of Informetrics*, 10 (2016), p. 224-237
2. Feng Feng et al. (2015). Visualization and quantitative study in bibliographic databases: A case in the field of university-industry cooperation, *Journal of Informetrics*, 9 (2015), p. 118-134
3. Guo Chen, Lu Xiao (2016). Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods, *Journal of Informetrics*, 10 (2016), p. 212-223
4. Ilia Reznik, Vladimir Shatalov, Hidden revolution of human priorities: An analysis of biographical data from Wikipedia, *Journal of Informetrics* 10 (2016) 124-131
5. Juan Gorraiz et al. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus, *Journal of Informetrics*, 10 (2016), p. 98-109.
6. Nicolai Mallig (2010). A relational database for bibliometric analysis, *Innovation Systems and Policy Analysis*, No. 22, ISSN 1612-1430, Karlsruhe.
7. Qi Yu et al. (2015). Tracing database usage: Detecting main paths in database link networks, *Journal of Informetrics*, 9 (2015), p. 1-15.

(Ngày Tòa soạn nhận được bài: 10-3-2017; Ngày phản biện đánh giá: 28-4-2017; Ngày chấp nhận đăng: 28-6-2017).