ỨNG DỤNG DỮ LIỆU LỚN TRONG CƠ QUAN THÔNG TIN-THƯ VIỆN

TS Ngô Thanh Thảo
Trường ĐH KHXH&NV- ĐHQG Tp. Hồ Chí Minh

Tóm tắt: *Bài viết giới thiệu khái quát về dữ liệu lớn, những thách thức, cơ hội và những vấn đề cần giải quyết khi ứng dụng dữ liệu lớn trong cơ quan thông tin-thư viện.*

Từ khóa: Dữ liệu lớn; ứng dụng dữ liệu lớn; cơ quan TT-TV.

Application of big data in information centers and libraries

Abstract: *The article introduces overview of big data, challenges, opportunities and issues to be solved when applying big data in information centers and libraries.*

Keywords: Big data; big data application; information centers and libraries.

Đặt vấn đề

Sự phát triển nhanh chóng của kỹ thuật thông tin số và công nghệ web dẫn đến sự gia tăng dữ liệu với quy mô vượt bậc trong nhiều lĩnh vực khác nhau. Từ những năm đầu của thế kỷ 21, nghiên cứu về dữ liệu lớn đã thu hút sự quan tâm đặc biệt của các nhà khoa học. Đến nay, dữ liệu lớn đã được ứng dụng thành công trong các loại hình tổ chức thuộc nhiều lĩnh vực khác nhau và đã đem lại nhiều cơ hội mới cho xã hội hiện đại. Ứng dụng công nghệ dữ liệu lớn nhằm tăng cường khả năng phục vụ người sử dụng cũng là vấn đề thu hút sự quan tâm của các nhà cung cấp dịch vụ thông tin hiện nay, trong đó có cơ quan thông tin-thư viện (CQTT-TV). Với công nghệ dữ liệu lớn, CQTT-TV có cơ hội quản trị, khai thác và sử dụng dữ liệu theo cách thức mới để tạo giá trị gia tăng cho sản phẩm, dịch vụ thông tin nhằm đáp ứng nhu cầu ngày càng đa dạng của người dùng tin.

Ứng dụng dữ liệu lớn có tầm quan trọng đặc biệt đối với CQTT-TV trong việc phân tích hành vi thông tin và nắm bắt nhu cầu tin của người dùng tin (NDT), trên cơ sở đó đưa ra biện pháp giải quyết các vấn đề quan trọng, như:

- xây dựng và khai thác hiệu quả nguồn tài nguyên thông tin;
- phát triển sản phẩm, dịch vụ thông tin theo hướng đa dạng hóa và cá nhân hóa để đáp ứng nhu cầu NDT;
- ứng dụng các phương tiện truyền thông thích hợp để tạo kênh tương tác hiệu quả giữa CQTT-TV và cộng đồng NDT;
- xây dựng và thực thi các chiến lược thích hợp để thu hút NDT,...

1. Khái quát về dữ liệu lớn

Hiện nay, có nhiều định nghĩa về dữ liệu lớn được đưa ra bởi các nhà nghiên cứu thuộc nhiều lĩnh vực khác nhau. Theo các nhà nghiên cứu của Viện Nghiên cứu Toàn cầu (McKinsey Global Institute), dữ liệu lớn là thuật ngữ dùng để chỉ tập hợp dữ liệu có khối lượng lớn đến mức vượt khả năng thu thập, lưu trữ, quản trị và phân tích của các công cụ và ứng dụng xử lý dữ liệu truyền thống [4]. Theo De Mauro, dữ liệu lớn là nguồn thông tin có đặc điểm là khối lượng lớn, tốc độ nhanh, đa dạng nên đòi hỏi phải có các công nghệ và phương pháp phân tích đặc trưng để khai thác được giá trị của nó [3].

Trong lĩnh vực thư viện - thông tin học cũng có nhiều định nghĩa khác nhau về dữ liệu lớn. Dựa trên kết quả phân tích, tổng hợp các định nghĩa về dữ liệu lớn được đề cập trong nhiều tài liệu khác nhau thuộc lĩnh vực này, các nhà nghiên cứu Phần Lan đã đưa ra định nghĩa "dữ liệu lớn là thuật ngữ dùng để chỉ tập hợp dữ liệu có khối lượng lớn, tốc độ gia tăng nhanh và đa dạng, do đó có thể làm phức tạp hóa các kỹ thuật xử lý dữ liệu nhưng đồng thời cũng thúc đẩy sự phát triển các giải pháp công nghệ" [7]. Mặc dù đưa ra những định nghĩa khác nhau, nhưng các nhà nghiên cứu lại có sự đồng thuận cao về đặc trưng của dữ liệu lớn, theo đó dữ liệu lớn được thể hiện bởi 3 đặc trưng cơ bản (gọi tắt là mô hình 3 V), như sau:

- Khối lượng (Volume): các tập dữ liệu của dữ liệu lớn có quy mô rất lớn so với dữ liệu truyền thống;

- Tốc độ (Velocity): khối lượng dữ liệu gia tăng nhanh chóng và tốc độ xử lý dữ liệu rất nhanh theo cơ chế xử lý thời gian thực.

- Đa dạng (Variety): dữ liệu đa dạng (có cấu trúc hoặc phi cấu trúc) và được thu thập từ nhiều nguồn khác nhau [8, 6].

2. Thách thức và cơ hội khi ứng dụng dữ liệu lớn trong cơ quan thông tin-thư viện

Các nhà nghiên cứu đã chứng minh rằng, dữ liệu trong CQTT-TV có các đặc trưng cơ bản của dữ liệu lớn là khối lượng, tốc độ và sự đa dạng. Vì vậy, có thể xem dữ liệu trong CQTT-TV là dữ liệu lớn. Dữ liệu lớn trong CQTT-TV được hình thành từ nhiều nguồn khác nhau, như:

- các bộ sưu tập tài liệu;

- dữ liệu về NDT;

- dữ liệu về các sản phẩm, dịch vụ thông tin- thư viện (SPDV TT-TV);

- dữ liệu về việc sử dụng các SPDV TT-TV; dữ liệu về sự tương tác giữa CQTT-TV với NDT qua các phương tiện truyền thông xã hội; …

Dữ liệu lớn được ứng dụng trong tất cả các hoạt động của CQTT-TV, bao gồm: thu thập, xử lý, tổ chức, lưu trữ và cung cấp thông tin [7]. Việc ứng dụng dữ liệu lớn có thể đem lại nhiều thách thức cũng như cơ hội cho CQTT-TV.

2.1. Thách thức

Khi ứng dụng dữ liệu lớn, CQTT-TV có thể phải đối mặt với những thách thức dưới đây [5,8].

2.1.1. Tính chính xác của dữ liệu

Như đã đề cập ở trên, dữ liệu trong CQTT-TVđa dạng về cấu trúc, bao gồm dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc. Điều này đòi hỏi phải có phương pháp thu thập và trình bày dữ liệu thích hợp để đảm bảo tính chính xác của dữ liệu. Tính chính xác của dữ liệu là yếu tố đặc biệt quan trọng đối với chất lượng của thông tin. Dữ liệu không chính xác sẽ làm giảm giá trị của dữ liệu gốc và làm tăng khối lượng công việc của khâu phân tích dữ liệu. Vì vậy, đảm bảo tính chính xác của dữ liệu là một trong những thách thức đối với CQTT-TV khi ứng dụng dữ liệu lớn.

2.1.2. Rút gọn và nén dữ liệu

Các CQTT-TV có rất nhiều dữ liệu, trong đó có cả những dữ liệu không hữu ích. Việc chọn lọc, rút gọn và nén dữ liệu rất cần thiết để đảm bảo giá trị của dữ liệu được lưu trữ không bị ảnh hưởng bởi những dữ liệu không hữu ích. Đồng thời, việc rút gọn và nén dữ liệu cũng có tác dụng làm giảm tải công việc của khâu phân tích dữ liệu. Mặc dù đây là công việc rất quan trọng nhưng trên thực tế hiện nay, các chuyên gia TT-TV còn thiếu các kỹ năng cần thiết để thực hiện việc rút gọn và nén dữ liệu. Và đây là một trong những thách thức mà các CQTT-TV phải vượt qua khi ứng dụng dữ liệu lớn.

2.1.3. Công nghệ và hệ thống xử lý dữ liệu lớn

Các hệ thống quản trị và phân tích dữ liệu được sử dụng trong các CQTT-TV hiện nay chỉ có thể áp dụng cho dữ liệu có cấu trúc và không thể đáp ứng được các yêu cầu kỹ thuật đối với việc thu thập, lưu trữ, xử lý và khai thác dữ liệu lớn. Công nghệ và hệ thống xử lý dữ liệu lớn có những ưu thế đặc biệt trong việc xử lý, phân tích dữ liệu bán cấu

trúc và phi cấu trúc. Tuy nhiên, CQTT-TV thường gặp hai trở ngại lớn khi ứng dụng công nghệ và hệ thống xử lý dữ liệu lớn, đó là chi phí cao và thiếu nguồn nhân lực có đủ khả năng vận hành hiệu quả công nghệ và hệ thống xử lý dữ liệu lớn. Vì vậy, công nghệ và hệ thống xử lý dữ liệu lớn thực sự là một thách thức lớn đối với CQTT-TV.

2.1.4. An toàn và bảo mật dữ liệu

Thông tin cá nhân của NDT thường được lưu trữ trong hệ thống thông tin của CQTT-TV. Do thiếu đội ngũ nhân viên có khả năng thực hiện tốt việc xử lý dữ liệu lớn nên hiện nay, nhiều CQTT-TV phải thuê các tổ chức chuyên nghiệp phân tích và xử lý dữ liệu của mình. Điều này có thể dẫn đến sự rò rỉ dữ liệu về NDT và làm gia tăng nguy cơ về an toàn dữ liệu. Vì vậy, đảm bảo an toàn và bảo mật dữ liệu là một trong những thách thức CQTT-TV phải đối mặt khi ứng dụng dữ liệu lớn.

2.2. Cơ hội

Bên cạnh những thách thức nêu trên, ứng dụng dữ liệu lớn cũng đem lại nhiều cơ hội phát triển cho CQTT-TV như sau [5,8]:

2.2.1. Làm phong phú CSDL

Khi ứng dụng dữ liệu lớn, dữ liệu trong CQTT-TV được tạo lập và trình bày với nhiều dạng thức khác nhau, như: văn bản, hình ảnh, âm thanh, video,… Những dữ liệu số này làm phong phú và đa dạng hóa CSDL, nhờ đó CQTT-TV có thể đáp ứng tốt hơn nhu cầu của NDT hiện tại và thu hút NDT tiềm năng.

2.2.2. Nâng cao chất lượng của nguồn nhân lực

Việc ứng dụng dữ liệu lớn đòi hỏi CQTT-TV phải có nguồn nhân lực đủ trình độ chuyên môn về quản lý và khai thác dữ liệu lớn. Để đáp ứng yêu cầu này, các CQTT-TV phải trang bị cho nhân viên những kiến thức và kỹ năng cần thiết cho việc thu thập, xử lý, lưu trữ, phân tích và khai thác dữ liệu lớn. Như vậy, ứng dụng dữ liệu lớn chính là cơ hội để CQTT-TV nâng cao chất lượng đội ngũ nhân viên của mình.

2.2.3. Phát triển dịch vụ mượn liên thư viện

Hiện nay, hầu hết các CQTT-TV đều phải đối mặt với vấn đề nan giải là không đủ kinh phí để phát triển nguồn tài nguyên thông tin nhằm đáp ứng nhu cầu ngày càng tăng của NDT. Chia sẻ nguồn tài nguyên thông tin qua dịch vụ mượn liên thư viện được xem như một giải pháp hữu hiệu để giải quyết vấn đề này. Đến nay, dịch vụ mượn liên thư viện là hoạt động chia sẻ các nguồn tài nguyên thông tin phổ biến nhất giữa các thư viện trên toàn cầu. Việc ứng dụng dữ liệu lớn sẽ giúp các CQTT-TV kịp thời nắm bắt nhu cầu của NDT và tăng cường chia sẻ thông tin về các nguồn tài liệu của các CQTT-TV, tạo điều kiện thuận lợi cho việc phát triển dịch vụ mượn liên thư viện.

2.2.4. Cung cấp các dịch vụ cá nhân hóa

Trong thời đại của dữ liệu lớn và internet, các dịch vụ thông tin cá nhân hóa có tầm quan trọng đặc biệt đối với sự phát triển của CQTT-TV. Ứng dụng công nghệ dữ liệu lớn trong việc thu thập, phân tích dữ liệu về các đặc điểm, sở thích và hành vi của NDT có thể cung cấp cho CQTT-TV thông tin hữu ích để phát triển các dịch vụ thông tin cá nhân hóa nhằm thỏa mãn tốt nhất nhu cầu tin của NDT. Bên cạnh đó, dựa trên kết quả phân tích dữ liệu về NDT, CQTT-TV có thể dự báo được nhu cầu tin và hành vi thông tin tiềm ẩn của NDT, từ đó có các giải pháp để thu hút NDT tiềm năng. Như vậy, ứng dụng dữ liệu lớn đem lại cơ hội phát triển các dịch vụ cá nhân hóa và thu hút NDT cho CQTT-TV.

3. Những vấn đề cần giải quyết khi ứng dụng dữ liệu lớn trong cơ quan thông tin-thư viện

Để ứng dụng hiệu quả dữ liệu lớn, CQTT-TV phải giải quyết nhiều vấn đề quan trọng, trong đó có các vấn đề liên quan đến nguồn nhân lực, nguồn tài nguyên thông tin, nâng cấp công nghệ, đổi mới dịch vụ và xây dựng hạ tầng cơ sở [5].

3.1. Nguồn nhân lực

Để quản trị và khai thác dữ liệu một cách hiệu quả, đội ngũ nhân viên của các CQTT-TV phải có kiến thức và kỹ năng cần thiết, như:

- kỹ năng thu thập, xử lý, tổ chức và bảo quản dữ liệu;

- kỹ năng lọc và nén dữ liệu;

- kỹ năng phân tích sâu dữ liệu;

- kỹ năng tạo thông tin hoặc kiến thức hữu ích từ dữ liệu lớn;

- kỹ năng giải quyết các vấn đề an toàn, bảo mật dữ liệu,...

Hiện nay, hầu hết các CQTT-TV đều thiếu nguồn nhân lực được trang bị đầy đủ những kỹ năng nói trên. Vì vậy, đào tạo nguồn nhân lực là yếu tố quan trọng, quyết định sự thành công khi ứng dụng dữ liệu lớn trong CQTT-TV. Trước mắt, các CQTT-TV có thể giải quyết vấn đề này theo nhiều cách khác nhau. Chẳng hạn, có thể chia nhân viên thành nhiều nhóm dựa trên lĩnh vực chuyên môn và kinh nghiệm thực tế để đào tạo theo những hướng khác nhau. Ví dụ, những nhân viên đã có hiểu biết về điện toán đám mây, internet vạn vật, dịch vụ di động phải được đào tạo theo hướng công nghệ. Còn những nhân viên có khả năng trong lĩnh vực tâm lý, marketing, quản lý thì có thể đào tạo theo hướng dịch vụ. Tuy nhiên, về lâu dài, việc đào tạo nguồn nhân lực có đủ khả năng ứng dụng hiệu quả dữ liệu lớn trong CQTT-TV phải được thực hiện một cách toàn diện bởi các cơ sở đào tạo chuyên ngành TT-TV. Chương trình đào tạo các chuyên gia TT-TV phải bao gồm những nội dung sau:

- Thu thập, tổ chức và bảo quản dữ liệu lớn: chương trình đào tạo phải trang bị cho người học các phương pháp và công cụ thu thập, đánh giá và chọn lọc các loại dữ liệu trong CQTT-TV, như: số liệu từ các cuộc khảo sát NDT, dữ liệu phân tích nguồn tài nguyên thông tin, kết quả thử nghiệm tính khả dụng của các SP-DV thông tin, dữ liệu về NDT, dữ liệu về mức độ thu hút NDT qua các phương tiện truyền thông,... Bên cạnh đó, người học phải được trang bị các kỹ năng tổ chức và bảo quản các loại dữ liệu khác nhau như văn bản, hình ảnh, số liệu thống kê,... cũng như kỹ năng xử lý các vấn đề về an toàn, bảo mật dữ liệu;

- Phân tích, khai thác dữ liệu lớn: người học phải được trang bị kiến thức và kỹ năng phân tích dữ liệu lớn trong các lĩnh vực như: tối ưu hóa kết quả tìm tin; phân tích và dự báo yêu cầu tin; lập kế hoạch phát triển nguồn tài nguyên thông tin; xây dựng chiến lược phát triển sản phẩm, dịch vụ thông tin; xây dựng chiến lược marketing,...

- Tạo lập, xử lý, quản trị, cung cấp nội dung: người học phải được trang bị kiến thức và kỹ năng tạo lập và cung cấp thông tin hữu ích cho NDT dựa trên dữ liệu lớn của CQTT-TV hoặc từ những nguồn khác;

- Nghiên cứu nhu cầu tin và thiết kế sản phẩm, dịch vụ đáp ứng nhu cầu tin;

- Nghiên cứu, thu thập, xử lý, tổ chức, khai thác, trình bày và phân phối thông tin;

- Tạo lập, chuyển giao và sử dụng thông tin;

- Quản trị các nguồn tài nguyên thông tin;

- Ứng dụng công nghệ thông tin và viễn thông để thiết kế, quảng bá và cung cấp các SPDV TT-TV;

- Quản lý CQTT-TV.

3.2. Nguồn tài nguyên thông tin

Để đáp ứng nhu cầu sử dụng tài liệu số ngày càng cao của NDT, CQTT-TV phải xây dựng nguồn tài nguyên số có nội dung phong phú và loại hình đa dạng. Việc xây dựng nguồn tài nguyên số phải dựa trên kết quả phân tích các loại dữ liệu khác nhau như: dữ liệu về sở thích, nhu cầu và thói quen dùng tin của NDT; dữ liệu về mức độ sử dụng các sản phẩm dịch vụ TT-TV,...

3.3. Nâng cấp công nghệ

Với trình độ công nghệ như hiện nay, các CQTT-TV rất khó có thể đáp ứng được các yêu cầu về điều kiện để thực hiện các công đoạn thu thập, xử lý, lưu trữ, phân tích và khai thác dữ liệu lớn. Vì vậy, CQTT-TV cần nâng cấp công nghệ nhằm đảm bảo điều kiện cần thiết để ứng dụng dữ liệu lớn. Chẳng hạn, CQTT-TV có thể sử dụng các công nghệ, như: NoSQL, PKI khi ứng dụng dữ liệu lớn. Do tính không đồng nhất của dữ liệu trong CQTT-TV nên NoSQL (Not Only SQL) là một lựa chọn hợp lý để xử lý, lưu trữ dữ liệu bán cấu trúc, phi cấu trúc cũng như

phát triển việc chia sẻ thông tin và hợp tác giữa các đơn vị.

Bên cạnh đó, CQTT-TV có thể ứng dụng PKI (Public Key Infrastructure - Hạ tầng khóa công khai) để đảm bảo sự an toàn, bảo mật dữ liệu. PKI là một công nghệ bảo mật mới bao gồm công nghệ khóa công khai và chiến lược bảo mật, chứng chỉ số và chứng thực số. Việc ứng dụng PKI rất hữu ích cho CQTT-TV trong việc bảo vệ bí mật cá nhân của NDT.

3.4. Đổi mới dịch vụ

Hành vi thông tin và cách thức sử dụng thông tin của NDT có sự thay đổi trong kỷ nguyên dữ liệu lớn nên các CQTT-TV phải tái định vị và đổi mới các dịch vụ của mình. Một trong những dịch vụ đổi mới là dịch vụ cung cấp thông tin cá nhân hóa dựa trên nền tảng cổng cá nhân. Với sự hỗ trợ của cổng cá nhân, các CQTT-TV có thể nhanh chóng thu thập thông tin hữu ích và gửi cho NDT một cách kịp thời. CQTT-TV cũng có thể cung cấp các dịch vụ cá nhân hóa qua nền tảng cổng cá nhân, như: đăng ký giữ trước tài liệu, cung cấp tài liệu qua e-mail, dịch vụ tư vấn,… Bên cạnh đó, CQTT-TV cũng cần phát triển các dịch vụ dành cho NDT đặc biệt, ví dụ như dịch vụ cung cấp tài liệu nhanh cho người khuyết tật. Với các dịch vụ được đổi mới, việc áp dụng dữ liệu lớn trong CQTT-TV sẽ thuận lợi và hiệu quả hơn.

3.5. Xây dựng hạ tầng cơ sở

Mặc dù hạ tầng cơ sở rất quan trọng đối với việc áp dụng dữ liệu lớn, nhưng hiện nay hầu hết các CQTT-TV đều thiếu kinh phí để xây dựng hạ tầng cơ sở. Để vượt qua trở ngại lớn này, CQTT-TVcó thể sử dụng các giải pháp như: tìm kiếm nguồn tài trợ từ các tổ chức hoặc các doanh nghiệp; phát triển các sản phẩm, dịch vụ thu phí; hợp tác và chia sẻ nguồn lực giữa các CQTT-TV,...

Kết luận

Ứng dụng dữ liệu lớn đem lại nhiều cơ hội cũng như thách thức cho các CQTT-TV. Để thực sự tận dụng được các cơ hội do công nghệ dữ liệu lớn đem lại, CQTT-TV phải giải quyết các vấn đề liên quan đến công nghệ, hạ tầng cơ sở, nguồn tài nguyên thông tin và đặc biệt là nguồn nhân lực. Trong điều kiện khó khăn về kinh phí hiện nay, CQTT-TV có thể giải quyết các vấn đề nêu trên dựa trên sự hợp tác và chia sẻ nguồn lực giữa các CQTT-TV và sự hỗ trợ tích cực từ các tổ chức liên quan như các cơ sở đào tạo chuyên ngành TT-TV, các nhà cung cấp thông tin, các nhà cung cấp giải pháp dữ liệu lớn, các tổ chức, doanh nghiệp…

TÀI LIỆU THAM KHẢO

1. Avinash S.S (2018). Big data: Application in Libraries, International Journal of Scientific Research in Multidisciplinary Studies, Vol.4, Issue 1, pp.22-23, January (2018). Truy cập từ http://isroset.org, ngày 02/04/2018.

2. Chen H., Doty P (2015). Library assessment and data analytics in the big data era: Practics and Policies. Truy cập từ https://onlinelibrary.wiley.com/doi/full/10…/pra2.2015.14505201002, ngày 02/04/2018.

3. De Mauro A (2016). A formal definition of big data based on its essential features Library Review, Vol. 65 Issue: 3, pp.122-135. Truy cập từ https://www.emeraldinsight.com/doi/pdfplus/10.1108/LR-06-2015-0061, ngày 12/04/2018.

4. James M (2011). Big data: The next frontier fo innovation, competition and productivity. Truy cập từ https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation, ngày 12/04/2018

5. Li J., Lu M (2017). Big data application framework and its feasibility analysis in library, Information Discovery and Delivery, Vol. 45 Issue: 4, pp.161-168, DOI: 10.1108/IDD-03-2017-0024.

6. Osman R.R (2017). The Evolution of data. From data to big data. Truy cập từ https://slaagc.org/…/The%20Evolution%20of%20Data.%20From%20D, ngày 20/04/2018.

7. Zhan M., Widen G (2017). Understanding big data in librarianship. Truy cập từ https://doi.org/10.1177%2F0961000617742451, ngày 20/04/2018.

8. Wang C (2016). Exposing Library data with big data technology: A Review. DOI: 10.1109/ICIS.2016.7550937.

(Ngày Tòa soạn nhận được bài: 15-12-2018; Ngày phản biện đánh giá: 20-02-2019; Ngày chấp nhận đăng: 15-3-2019).