TỔNG QUAN VỀ DỮ LIỆU LỚN (BIGDATA)

KS. Nguyễn Công Hoan
Trung Tâm Thông tin khoa học Thống kê, Viện KHTK

Trước đây, chúng ta mới chỉ biết đến dữ liệu có cấu trúc (structure data), ngày nay, với sự kết hợp của dữ liệu và internet, đã xuất hiện một dạng khác của dữ liệu - Big data (dịch là "dữ liệu lớn"). Dữ liệu này có thể từ các nguồn như: hồ sơ hành chính,giao dịch điện tử, dòng trạng thái (status), chia sẻ hình ảnh, bình luận, tin nhắn... của chính chúng ta, nói cách khác chúng là dữ liệu được sản sinh qua quá trình chia sẻ thông tin trực tuyến liên tục của người sử dụng. Để cung cấp cái nhìn tổng quan, bài viết này giới thiệu tóm tắt những nét chính về dữ liệu lớn cũng như những cơ hội và thách thức mà dữ liệu lớn mang lại.

1. Khái niệm, đặc trưng của dữ liệu lớn và sự khác biệt với dữ liệu truyền thống

1.1. Khái niệm về dữ liệu lớn

- *Theo wikipedia*: Dữ liệu lớn (Big data) là một thuật ngữ chỉ bộ dữ liệu lớn hoặc phức tạp mà các phương pháp truyền thống không đủ các ứng dụng để xử lý dữ liệu này.

- *Theo Gartner*: Dữ liệu lớn là những nguồn thông tin có đặc điểm chung khối lượng lớn, tốc độ nhanh và dữ liệu định dạng dưới nhiều hình thức khác nhau, do đó muốn khai thác được đòi hỏi phải có hình thức xử lý mới để đưa ra quyết định, khám phá và tối ưu hóa quy trình.

1.2. Nguồn hình thành và phương pháp khai thác, quản lý dữ liệu lớn

Qua thống kê và tổng hợp, dữ liệu lớn được hình thành chủ yếu từ 6 nguồn: (1) Dữ liệu hành chính (phát sinh từ chương trình của một tổ chức, có thể là chính phủ hay phi chính phủ). Ví dụ, hồ sơ y tế điện tử ở bệnh viện, hồ sơ bảo hiểm, hồ sơ ngân hàng...; (2) Dữ liệu từ hoạt động thương mại (phát sinh từ các giao dịch giữa hai thực thể). Ví dụ, các giao dịch thẻ tín dụng, giao dịch trên mạng, bao gồm cả các giao dịch từ các thiết bị di động; (3) Dữ liệu từ các thiết bị cảm biến như thiết bị chụp hình ảnh vệ tinh, cảm biến đường, cảm biến khí hậu; (4) Dữ liệu từ các thiết bị theo dõi, ví dụ theo dõi dữ liệu từ điện thoại di động, GPS; (5) Dữ liệu từ các hành vi, ví dụ như tìm kiếm trực tuyến (tìm kiếm sản phẩm, dịch vụ hay thông tin khác), đọc các trang mạng trực tuyến...; (6) Dữ liệu từ các thông tin về ý kiến, quan điểm của các cá nhân, tổ chức, trên các phương tiện thông tin xã hội.

Phương pháp khai thác và quản lý dữ liệu lớn hiện nay được thiết kế phù hợp dựa theo các nguồn hình thành dữ liệu lớn. Mỗi nguồn dữ liệu lớn khác nhau sẽ có phương pháp khai thác và quản lý dữ liệu lớn khác nhau. Tuy nhiên, hiện nay phần lớn các tổ chức trên thế giới đều dùng Hadoop ecosystem là giải pháp tối ưu để khai thác và quản lý dữ liệu lớn.

1.3. Đặc trưng 5V của dữ liệu lớn

Dữ liệu lớn có 5 đặc trưng cơ bản như sau (mô hình 5V[9]):

(1) Khối lượng dữ liệu (Volume)

Đây là đặc điểm tiêu biểu nhất của dữ liệu lớn, khối lượng dữ liệu rất lớn. Kích cỡ của Big data đang từng ngày tăng lên, và tính đến năm 2012 thì nó có thể nằm trong khoảng vài chục terabyte cho đến nhiều petabyte (1 petabyte = 1024 terabyte) chỉ cho một tập hợp dữ liệu. Dữ liệu truyền thống có thể lưu trữ trên các thiết bị đĩa mềm, đĩa cứng. Nhưng với dữ liệu lớn chúng ta sẽ sử dụng công nghệ "đám mây" mới đáp ứng khả năng lưu trữ được dữ liệu lớn.

(2) Tốc độ (Velocity)

Tốc độ có thể hiểu theo 2 khía cạnh: (a) Khối lượng dữ liệu gia tăng rất nhanh (mỗi giây có tới 72.9 triệu các yêu cầu truy cập tìm kiếm trên web bán hàng của Amazon); (b) Xử lý dữ liệu nhanh ở mức thời gian thực (real-time), có nghĩa dữ liệu được xử lý ngay tức thời ngay sau khi chúng phát sinh (tính đến bằng mili giây). Các ứng dụng phổ biến trên lĩnh vực Internet, Tài chính, Ngân hàng, Hàng không, Quân sự, Y tế – Sức khỏe như hiện nay phần lớn dữ liệu lớn được xử lý real-time. Công nghệ xử lý dữ liệu lớn ngày nay đã cho phép chúng ta xử lý tức thì trước khi chúng được lưu trữ vào cơ sở dữ liệu.

(3) Đa dạng (Variety)

Đối với dữ liệu truyền thống chúng ta hay nói đến dữ liệu có cấu trúc, thì ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc (tài liệu, blog, hình ảnh, vi deo, bài hát, dữ liệu từ thiết bị cảm biến vật lý, thiết bị chăm sóc sức khỏe…). Big data cho phép liên kết và phân tích nhiều dạng dữ liệu khác nhau. Ví dụ, với các bình luận của một nhóm người dùng nào đó trên Facebook với thông tin video được chia sẻ từ Youtube và Twitter.

---

[9] *Mô hình 5Vs do Gartner xây dựng. Gartner là công ty nghiên cứu và tư vấn về công nghệ thông tin hàng đầu thế giới do một người Mỹ tên là Gideon Gartner sáng lập năm 1979. 5Vs: Khối lượng dữ liệu (Volume); Tốc độ (Velocity); Giá trị (Value); Độ tin cậy/chính xác (Veracity); Đa dạng (Variety).*

(4) Độ tin cậy/chính xác (Veracity)

Một trong những tính chất phức tạp nhất của Dữ liệu lớn là độ tin cậy/chính xác của dữ liệu. Với xu hướng phương tiện truyền thông xã hội (Social Media) và mạng xã hội (Social Network) ngày nay và sự gia tăng mạnh mẽ tính tương tác và chia sẻ của người dùng Mobile làm cho bức tranh xác định về độ tin cậy & chính xác của dữ liệu ngày một khó khăn hơn. Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là tính chất quan trọng của Big data.

(5) Giá trị (Value)

Giá trị là đặc điểm quan trọng nhất của dữ liệu lớn, vì khi bắt đầu triển khai xây dựng dữ liệu lớn thì việc đầu tiên chúng ta cần phải làm đó là xác định được giá trị của thông tin mang lại như thế nào, khi đó chúng ta mới có quyết định có nên triển khai dữ liệu lớn hay không. Nếu chúng ta có dữ liệu lớn mà chỉ nhận được 1% lợi ích từ nó, thì không nên đầu tư phát triển dữ liệu lớn. Kết quả dự báo chính xác thể hiện rõ nét nhất về giá trị của dữ liệu lớn mang lại. Ví dụ, từ khối dữ liệu phát sinh trong quá trình khám, chữa bệnh sẽ giúp dự báo về sức khỏe được chính xác hơn, sẽ giảm được chi phí điều trị và các chi phí liên quan đến y tế.

1.4. Sự khác biệt giữa dữ liệu lớn với dữ liệu truyền thống

Dữ liệu lớn khác với dữ liệu truyền thống (ví dụ, kho dữ liệu - Data Warehouse) ở 4 điểm cơ bản: Dữ liệu đa dạng hơn; lưu trữ dữ liệu lớn hơn; truy vấn dữ liệu nhanh hơn; độ chính xác cao hơn.

(1) Dữ liệu đa dạng hơn: Khi khai thác dữ liệu truyền thống (dữ liệu có cấu trúc), chúng ta thường phải trả lời các câu hỏi: Dữ liệu lấy ra kiểu gì? định dạng dữ liệu như thế nào? Đối với dữ liệu lớn, không phải trả lời các câu hỏi trên. Hay nói khác, khi khai thác, phân tích dữ liệu lớn chúng ta không cần quan tâm đến kiểu dữ liệu và định dạng của chúng; điều quan tâm là giá trị mà dữ liệu mang lại có đáp ứng được cho công việc hiện tại và tương lai hay không.

(2) Lưu trữ dữ liệu lớn hơn: Lưu trữ dữ liệu truyền thống vô cùng phức tạp và luôn đặt ra câu hỏi lưu như thế nào? dung lượng kho lưu trữ bao nhiêu là đủ? gắn kèm với câu hỏi đó là chi phí đầu tư tương ứng. Công nghệ lưu trữ dữ liệu lớn hiện nay đã phần nào có thể giải quyết được vấn đề trên nhờ những công nghệ lưu trữ đám mây, phân phối lưu trữ dữ liệu phân tán và có thể kết hợp các dữ liệu phân tán lại với nhau một cách chính xác và xử lý nhanh trong thời gian thực.

(3) Truy vấn dữ liệu nhanh hơn: Dữ liệu lớn được cập nhật liên tục, trong khi đó kho dữ liệu truyền thống thì lâu lâu mới được cập nhật và trong tình trạng không theo dõi thường xuyên gây ra tình trạng lỗi cấu trúc truy vấn dẫn đến không tìm kiếm được thông tin đáp ứng theo yêu cầu.

(4) Độ chính xác cao hơn: Dữ liệu lớn khi đưa vào sử dụng thường được kiểm định lại dữ liệu với những điều kiện chặt chẽ, số lượng thông tin được kiểm tra thông thường rất lớn, và đảm bảo về nguồn lấy dữ liệu không có sự tác động của con người vào thay đổi số liệu thu thập.

2. Bức tranh tổng thể ứng dụng dữ liệu lớn

Dữ liệu lớn đã được ứng dụng trong nhiều lĩnh vực như: hoạt động chính trị; giao thông; y tế; thể thao; tài chính; thương mại; thống kê... dưới đây là một số ví dụ về ứng dụng dữ liệu lớn.

2.1. Ứng dụng dữ liệu lớn trong hoạt động chính trị

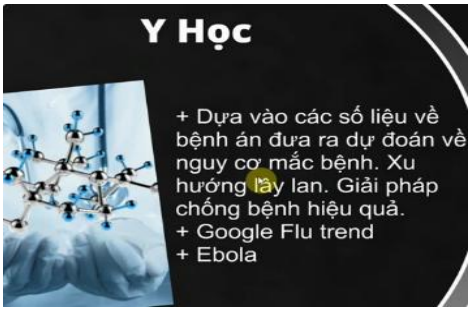
Hình bên cho thấy Tổng thống Mỹ Obama đã sử dụng dữ liệu dữ liệu lớn để phục vụ cho cuộc tranh cử Tổng thống của mình. Ông xây dựng một đội ngũ nhân viên chuyên đi thu thập thông tin và phân tích dữ liệu thu được trong dự án triển khai về dữ liệu lớn. Đội ngũ nhân viên này thu thập tất cả thông tin về người dân ở các khu vực, sau đó phân tích và chỉ ra một số thông tin quan trọng về người dân Mỹ như: Thích đọc sách gì, thích mua loại thuốc gì, thích sử dụng phương tiện gì… Thậm chí còn biết được cả thông tin về mẹ của cử tri đó đã bỏ phiếu tín nhiệm ai ở lần bầu cử trước. Trên cơ sở những thông tin này, Tổng thống Obama đã đưa ra kế hoạch vận động phù hợp, giúp ông tái đắc cử Tổng thống nước Mỹ lần thứ 2.

Ngoài ra một số ứng dụng khác trong lĩnh vực chính trị mà dữ liệu lớn được áp dụng như: Hệ thống chính phủ điện tử; phân tích quy định và việc tuân thủ quy định; phân tích, giám sát, theo dõi và phát hiện gian lận, mối đe dọa, an ninh mạng.

2.2. Ứng dụng dữ liệu lớn trong giao thông

Sử dụng số liệu CDR trong quá khứ để ước lượng các dòng giao thông trong thành phố vào các giờ cao điểm, từ đó có những kế hoạch phân luồng giao thông chi tiết, hợp lý giúp giảm thiểu kẹt xe. Ngoài ra còn đưa ra thông tin cho người tham gia giao thông được biết nếu muốn đi từ nơi này đến nơi khác thì nên đi vào giờ nào để tránh kẹt xe, hoặc đi đường nào là ngắn nhất, v.v... Ngoài ra, dữ liệu lớn còn giúp phân tích định vị người dùng thiết bị di động, ghi nhận chi tiết cuộc gọi trong thời gian thực; và giảm thiểu tình trạng ùn tắc giao thông.

2.3. Ứng dụng dữ liệu lớn trong y tế



Trong y học các bác sĩ dựa vào số liệu trong các bệnh án để đưa ra dự đoán về nguy cơ mắc bệnh. Đồng thời cũng đưa ra được xu hướng lây lan của bệnh. Ví dụ, ứng dụng Google Flu Trend là một trong những ứng dụng thành công của Google ứng dụng này dựa trên từ khóa tìm kiếm ở một khu vực nào đó, sau đó bộ máy phân tích của Google sẽ phân tích và đối chiếu kết quả tìm kiếm đó, sau cùng là đưa ra dự báo về xu hướng dịch cúm tại khu vực đó. Qua đó cho biết tình hình cúm tại khu vực đó sẽ diễn ra như thế nào để đưa ra các giải pháp phòng tránh. Những kết quả mà Google Flu Trend đưa ra, hoàn toàn phù hợp với báo cáo của Tổ chức Y tế Thế giới WHO về tình hình bệnh cúm tại các khu vực đó.

2.4. Ứng dụng dữ liệu lớn trong thể thao

Phân tích mô hình hệ thống cấu trúc sơ đồ chiến thuật của đội tuyển Đức (hình bên) đã đưa ra những điểm bất hợp lý trong cấu trúc của đội tuyển Đức, từ đó giúp cho đội tuyển Đức khắc phục được điểm yếu và đã dành được World cup 2014.

2.5. Ứng dụng dữ liệu lớn trong tài chính

Từ những dữ liệu chính xác, kịp thời thu thập được thông qua các giao dịch của khách hàng, tiến hành phân tích, xếp hạng và quản lý các rủi ro trong đầu tư tài chính, tín dụng.

2.6. Ứng dụng dữ liệu lớn trong thương mại

Trong thương mại dữ liệu lớn giúp cho chúng ta thực hiện được một số công việc sau: Phân khúc thị trường và khách hàng; phân tích hành vi khách hàng tại cửa hàng; tiếp thị trên nền tảng định vị; phân tích tiếp thị chéo kênh, tiếp thị đa kênh; quản lý các chiến dịch tiếp thị và khách hàng thân thiết; So sánh giá; Phân tích và quản lý chuỗi cung ứng; Phân tích hành vi, thói quen người tiêu dùng.

2.7. Ứng dụng dữ liệu lớn trong thống kê

Nhận thấy những lợi ích to lớn và thách thức của Bigdata đối với thống kê nhà nước, Ủy ban Thống kê Liên hợp quốc cũng như các tổ chức thống kê khu vực và Cơ quan thống kê quốc gia của nhiều nước đã triển khai hàng loạt các hoạt động về Bigdata như: Hàn Quốc sử dụng ảnh vệ tinh để thống kê nông nghiệp và một số lĩnhvực khác; Australia sử dụng ảnh vệ tinh để thống kê diện tích đất nông nghiệp và năng suất; Italia sử dụng dữ liệu điện thoại di động để thống kê di cư; Bhutan dùng

thiết bị di động để tính toán chỉ số giá tiêu dùng; Estonia dùng điện thoại di động định vị vệ tinh để thống kê du lịch; EuroStat sử dụng dữ liệu về sử dụng điện thoại di động để thống kê du lịch[10].

3. Cơ hội và thách thức khi ứng dụng dữ liệu lớn trong thống kê Nhà nước

3.1 Cơ hội

(1) Tiếp cận và nghiên cứu về dữ liệu lớn sẽ giúp cho chúng ta có thêm phương án giải quyết, xử lý và đối phó với những thách thức đối sản xuất số liệu thống kê nhà nước trong hiện tại và tương lai. Những nghiên cứu thực nghiệm cần phải được tiến hành để khám phá những ứng dụng tiềm năng của dữ liệu lớn trong số liệu thống kê nhà nước, và nghiên cứu thực nghiệm đó phải là một phần trong Quy trình sản xuất số liệu thống kê.

(2) Nghiên cứu về dữ liệu lớn cần phải có cơ sở hạ tầng công nghệ thông tin hiện đại, đáp ứng các yêu cầu xử lý khối lượng lớn dữ liệu và nhanh, đồng thời có thể tập hợp dữ liệu từ nhiều nguồn khác nhau. Thực hiện được điều này chúng ta có được đội ngũ nguồn nhân lực về quản lý và khai thác Big data vững vàng về chuyên môn và được trải qua kinh nghiệm thực tế.

(3) Tiếp cận và nghiên cứu về dữ liệu lớn sẽ giúp chúng ta có được những văn bản pháp lý bổ sung có thể giúp cho Cơ quan thống kê nhà nước có điều kiện để thực hiện được khai thác dữ liệu thông qua hồ sơ hành chính, ngoài ra dữ liệu cũng được bảo đảm và giữ bí mật nhờ những văn bản pháp lý bổ sung này.

(4) Sử dụng dữ liệu lớn đem lại niềm tin của cộng đồng với thống kê nhà nước do quá trình trình sản xuất số liệu thống kê nhà nước với dữ liệu lớn hoàn toàn không có sự tác động chủ ý của con người.

3.2 Thách thức

(1) Tài chính

Nhiều đơn vị, tổ chức không đo lường được vấn đề sẽ phát sinh trong quá trình triển khai thực hiện, dự toán kinh phí chưa chính xác, do vậy dự án không thực hiện được. Để triển khai được thành công, yếu tố tài chính có ý nghĩa rất quan trọng, một số tập đoàn thương mại lớn có tiềm lực tài chính vững chắc đã xây dựng thuận lợi hệ thống dữ liệu Big data như IBM, website bán hàng thương mại điện tử Amazon...

(2) Chính sách, quy định Luật pháp về truy cập và sử dụng dữ liệu

Việc sử dụng và khai thác dữ liệu lớn phụ thuộc vào Luật quy định của mỗi quốc gia.

[10] *Xem Báo cáo "Thống kê Nhà nước với Big data: Kinh nghiệm quốc tế và định hướng của Thống kê Việt Nam"*

Ví dụ: ở Canada người dùng có thể được tiếp cận dữ liệu từ cả hai tổ chức chính phủ và phi chính phủ, nhưng ở những nước khác như Ireland thì phải được sự cho phép từ các cơ quan chính phủ. Điều này có thể dẫn đến những hạn chế để truy cập vào một số loại dữ liệu lớn.

(3) Trình độ khai thác và quản lý dữ liệu

Do Luật quy định sử dụng và khai thác ở mỗi quốc gia là khác nhau nên cách quản lý là cũng khác nhau tuy nhiên, một vấn đề liên quan đến quản lý thông tin hiện nay là nguồn nhân lực. Khoa học dữ liệu lớn đang phát triển mạnh trong những tổ chức tư nhân, trong khi đó bộ phận này chưa được liên kết với những tổ chức của chính phủ một cách chặt chẽ dẫn đến việc quản lý vẫn còn nhiều vướng mắc.

(4) Hạ tầng công nghệ thông tin

Cần phải cải thiện tốc độ dữ liệu truy cập vào các dữ liệu hành chính nghĩa là có thể sử dụng giao diện ứng dụng của Chương trình chuyên sâu tiêu chuẩn (API) để truy cập dữ liệu. Bằng cách này, nó có thể kết nối các ứng dụng cho dữ liệu thu về và xử lý dữ liệu trực tiếp với dữ liệu hành chính. Ngoài ra hệ thống khai thác dữ liệu lớn cũng cần phải được tính toán để có thể kết nối vào được kho cơ sở dữ liệu truyền thống, đó cũng là một trong những thách thức lớn cần được giải quyết.

Tóm lại

Trong bài nghiên cứu trên chúng tôi đã đưa ra được những thông tin cơ bản về Big data, những lợi ích mà Big data mang lại. Bên cạnh đó cũng chỉ ra những thách thức khi triển khai áp dụng khai thác Big data.

Điều quan trọng nhất trong báo cáo này đã đưa ra những ưu điểm của Big data đó là cung cấpthông tin để chung ta xử lý được tình huống nhanh nhất, chính xác nhất và giá trị của Big data mang lại luôn có tính định hướng đến tương lai?; Giải đáp những câu hỏi tại sao việc ấy lại xảy ra?; Sau chuyện đó thì điều gì sẽ sảy ra? và chúng ta nên ứng phó như thế nào trong hoàn cảnh đó?

Tài liệu tham khảo

1. Tài liệu cơ hội và thách thức với bigdata –E của Liên hợp quốc: http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf

2. Báo cáo Hội thảo về tương lai của Thống kê học London: https://statistics.stanford.edu/statistics-and-science-london-workshop-report

3. Tài liệu về các khái niệm và đặc trưng của Big data: https://viblo.asia/dovv/posts/3OEqGjWwv9bL