VỚI SINH VIÊN TỐT NGHIỆP NGÀY NAY, CHỈ MỘT TỪ: THỐNG KÊ

Steve Lohr

Steve Lohr, phóng viên chuyên mục Công nghệ, kinh doanh và kinh tế của tờ New York Time, ngày 6 tháng 9 năm 2009 có bài viết về vai trò quan trọng của các chuyên gia phân tích dữ liệu thống kê và máy tính trong xã hội ngày nay. Dưới đây là nội dung của bài báo.

Tại Đại học Harvard, Carrie Grimes theo học chuyên ngành nhân chủng học và khảo cổ học và đã mạo hiểm đến những nơi như Honduras, ở đây cô nghiên cứu các mô hình định cư của người Maya bằng cách lập bản đồ hiện vật đã tìm được. Nhưng cô đã rút ra là máy tính và toán học là một phần của ngành khảo cổ.

Cô nói "Mọi người nghĩ về lĩnh vực khảo cổ học giống như nhân vật Indiana Jones trong bộ phim cùng tên, nhưng phần lớn những gì bạn thực sự làm là phân tích dữ liệu".

Bây giờ, Grimes thực hiện một loại công việc khai quật khác. Cô làm việc tại Google, nơi cô sử dụng phân tích thống kê hàng đống dữ liệu để đưa ra biện pháp giúp cải thiện công cụ tìm kiếm của công ty.

Cô Grimes là một nhà thống kê trong kỷ nguyên Internet, một trong số nhiều người đang thay đổi hình ảnh của nghề nghiệp chuyên môn là một người say mê nghiên cứu những con số. Nhu cầu đối với công việc thống kê này đang gia tăng.

Ông Hal Varian, giám đốc kinh tế tại Google nói: "Tôi thường nói rằng công việc hấp dẫn trong vòng 10 năm tới sẽ là thống kê, và tôi không đùa".

Vị thế gia tăng của các chuyên gia thống kê, những người có thể kiếm được 125.000 đô la tại các công ty hàng đầu trong năm đầu tiên của họ sau khi có bằng tiến sĩ, là hệ quả của sự bùng nổ dữ liệu số trong thời gian gần đây. Trong nhiều lĩnh vực khác nhau, máy tính và web đang tạo ra các lĩnh vực mới để thăm dò dữ liệu - tín hiệu cảm biến, băng ghi âm theo dõi, trò chuyện trên mạng xã hội, các tài liệu công và nhiều thứ khác nữa. Và theo dự báo của công ty nghiên cứu IDC thì sự gia tăng dữ liệu số sẽ tăng tốc, tăng gấp 5 lần vào năm 2012.

Tuy nhiên, dữ liệu chỉ là nguyên liệu thô của tri thức. "Chúng ta đang nhanh chóng bước vào một thế giới nơi mà tất cả mọi thứ có thể được theo dõi và đo lường," ông Erik Brynjolfsson, nhà kinh tế, giám đốc Trung tâm doanh nghiệp kỹ thuật số thuộc Viện Công nghệ Massachusetts nói. "Nhưng vấn đề lớn phải giải quyết sẽ là khả năng của con người đối với việc sử dụng, phân tích và tạo ra ý nghĩa của dữ liệu."

Thế hệ chuyên gia thống kê mới phải giải quyết vấn đề đó. Họ sử dụng máy tính mạnh và các mô hình toán học phức tạp. Các ứng dụng cũng rất đa dạng như cải thiện công cụ tìm kiếm Internet và quảng cáo trực tuyến, chọn lọc thông tin về sắp xếp trật tự gien trong nghiên cứu ung thư, phân tích các dữ liệu cảm biến và định vị để tối ưu hóa quy trình vận chuyển thực phẩm.

Thậm chí gần đây đã kết thúc cuộc thi Netflix (Netflix là thương hiệu hàng đầu trong lĩnh vực dịch vụ phim ảnh trực tuyến ở Mỹ) sẵn sàng cấp một triệu đô

la cho bất cứ ai có khuyến nghị để cải thiện đáng kể hệ thống lọc thông tin về phim ảnh của công ty, là một trận chiến bắt đầu với vũ khí thống kê hiện đại.

Tuy nhiên, các nhà thống kê chỉ là một phần nhỏ của lực lượng các chuyên gia sử dụng các kỹ thuật thống kê hiện đại để phân tích dữ liệu. Các chuyên gia cho rằng kỹ năng sử dụng máy tính và số học, có ý nghĩa quan trọng hơn nhiều so với bằng cấp. Do đó, các chuyên gia phân tích dữ liệu mới nắm vững các kiến thức cơ bản như kinh tế, khoa học máy tính và toán học.

Họ chắc chắn được Nhà Trắng chào đón vào những ngày này. “Mạnh mẽ, dữ liệu khách quan là bước đầu tiên hướng tới giải quyết các nhu cầu kinh tế dài hạn và các ưu tiên chính sách quan trọng của chúng tôi,” Peter R. Orszag, giám đốc Văn phòng Quản lý và Ngân sách, đã tuyên bố trong một bài phát biểu vào tháng Năm. Sau ngày đó, ông Orszag (được thừa nhận là người am tường về kinh tế, chính trị, ngoại giao) đã thú nhận trên một blog tham gia nói chuyện về vai trò quan trọng của thống kê là một chủ đề “gần trái tim tôi.”

IBM nhận thấy cơ hội trong các dịch vụ săn tìm dữ liệu, đã lập ra nhóm Phân tích kinh doanh và Dịch vụ tối ưu hóa vào tháng Tư. Nhóm sẽ đặt quan hệ xin ý kiến chuyên môn của hơn 200 nhà toán học, nhà thống kê và các nhà phân tích dữ liệu khác trong phòng thí nghiệm nghiên cứu của nó - nhưng số lượng không đủ. IBM lên kế hoạch đào tạo lại hoặc thuê hơn 4.000 chuyên gia phân tích cho toàn công ty.

Dấu hiệu khác về sự quan tâm ngày càng tăng trong lĩnh vực này, theo Hiệp hội thống kê Mỹ ước tính có khoảng 6.400 người tham dự hội nghị chuyên môn thống kê thường niên ở Washington tuần này, tăng từ khoảng 5.400 người trong những năm gần đây. Những người tham dự hội nghị, nam và nữ, trẻ và già, trông giống như bất kỳ đám đông khách du lịch nào tại thủ đô. Tuy nhiên, những trao đổi say mê của họ được lấp đầy các từ về ngẫu nhiên, các tham số, hồi quy và bộ các tập hợp con trong mẫu thống kê. Sự đột biến về dữ liệu đang làm tăng giá trị nghề nghiệp mà giải quyết công việc truyền thống có thể ít rõ ràng và công việc ít sinh lợi, như tìm ra các số liệu về mức tuổi thọ bình quân cho các công ty bảo hiểm.

Cô Grimes, 32 tuổi, nhận học vị tiến sĩ thống kê từ Stanford vào năm 2003 và đã gia nhập Google vào cuối năm đó. Cô hiện đang là một trong số nhiều nhà thống kê của một nhóm gồm 250 nhà phân tích dữ liệu. Cô sử dụng mô hình thống kê để cải thiện công nghệ tìm kiếm của công ty.

Ví dụ, cô Grimes tiếp tục phát triển thuật toán liên quan đến tìm kiếm tự động của Google, cảnh báo web theo hướng liên tục cập nhật chỉ số tìm kiếm.

Cô Grimes đã giải thích mục tiêu là để tạo ra lợi ích rất nhỏ về hiệu quả sử dụng máy tính và mạng. “Ngay một sự cải tiến một hoặc hai phần trăm có thể là rất lớn, khi bạn thực hiện hàng triệu và hàng tỷ lần chúng ta làm các công việc tại Google," cô nói.

Đó là quy mô của bộ dữ liệu trên web sẽ mở ra thế giới mới của sự khám phá. Theo truyền thống, khoa học xã hội theo dõi hành vi của người dân bằng cách phỏng vấn hoặc điều tra họ. “Tuy nhiên, web cung cấp nguồn tài nguyên đáng ngạc nhiên này để quan sát hàng triệu người tương tác như thế nào,” ông Jon Kleinberg, một nhà khoa học máy tính và nhà nghiên cứu mạng xã hội tại Cornell đã nói.

Ví dụ, trong nghiên cứu mới công bố, ông Kleinberg và hai đồng nghiệp đã theo dõi dòng ý tưởng trên không gian ảo. Họ theo dõi 1,6 triệu trang web tin tức và blog trong suốt chiến dịch vận động

tranh cử tổng thống 2008, sử dụng các thuật toán để lọc các cụm từ liên kết với các chủ đề tin tức như "che đậy". Các nhà nghiên cứu Cornell thấy rằng, thông thường, dẫn đầu là phương tiện truyền thống và sau đó là các blog, thường sau 2,5 giờ. Tuy nhiên, có một số ít blog nhanh nhất đối với việc trích dẫn sau khi đã nhận được sự quan tâm từ nhiều người.

Các chuyên gia cảnh báo, những mạch nhỏ phóng phú của dữ liệu web, có sự nguy hiểm của nó. Khối lượng tuyệt đối của nó có thể dễ dàng áp đảo các mô hình thống kê. Các nhà thống kê cũng cảnh báo rằng mối tương quan mạnh của dữ liệu không nhất thiết chứng tỏ một sự kết nối nhân - quả.

Ví dụ, vào cuối những năm 1940, trước khi có vaxin bại liệt, các chuyên gia y tế công cộng ở Mỹ đã ghi nhận rằng số ca mắc bại liệt tăng lên cùng với mức tiêu thụ kem và nước giải khát, theo David Alan Grier, một sử gia và nhà thống kê tại đại học George Washington. Việc loại bỏ kem và nước giải khát thậm chí còn được khuyến cáo như là một phần của chế độ ăn uống chống lại bệnh bại liệt. Hóa ra là sự bùng nổ của bệnh bại liệt phổ biến nhất trong những tháng nóng của mùa hè, khi mọi người ăn kem nhiều hơn, chỉ biểu thị một sự liên đới, ông Grier nói. Nếu việc khai thác dữ liệu cường điệu các vấn đề đã có từ lâu về thống kê, thì nó cũng mở ra một lĩnh vực mới.

"Điều quan trọng là để máy tính làm những gì chúng làm tốt, tìm ra trong những bộ dữ liệu rất lớn những điều khác thường", ông Daniel Gruhl, một nhà nghiên cứu IBM, công việc gần đây của ông trong đó có khai thác dữ liệu y tế để cải thiện việc điều trị. "Và điều đó làm cho con người có khả năng làm tốt nhất - lý giải những điều khác thường đó".

Nguyễn Thái Hà (dịch)

For Today's Graduate, Just One Word: Statistics
http://topics.nytimes.com/top/reference/timestopics/people/l/steve_lohr/....