

ỨNG DỤNG VÀ KHAI THÁC NGUỒN DỮ LIỆU LỚN TRONG THỐNG KÊ

Nguyễn Thị Nhung

Tóm tắt

Ngày nay, nhờ sự phát triển của internet cùng mạng lưới công nghệ thông tin (CNTT) nên dữ liệu lớn (DLL-Big data) là một sự tồn tại tất yếu trong xã hội. Dữ liệu đã trở thành yếu tố dẫn dắt trong nhiều hoạt động của công nghệ. Những cải tiến mới liên tục đưa ra những khám phá về dữ liệu, về cách chúng ta sử dụng và phân tích dữ liệu. Bài viết tập trung nghiên cứu về một số nội dung cơ bản về dữ liệu, dữ liệu lớn và với nguồn dữ liệu lớn, nguồn dữ liệu có nhiều tính năng nổi trội so với nguồn dữ liệu truyền thống thì công tác thống kê phân tích, xử lý như thế nào. Qua đó, nghiên cứu đề xuất một số giải pháp góp phần đẩy mạnh công tác ứng dụng và khai thác nguồn dữ liệu lớn trong thống kê trong thời gian tới

Từ khóa: Dữ liệu, dữ liệu lớn, thống kê.

APPLICATIONS AND EXPLOITATION OF BIG DATA RESOURCES IN STATISTICS

Abstract

Currently, the existence of big data has become an inevitable phenomenon in society due to the rapid development of the internet and information technology (IT) networks. Data has emerged as a significant driving force behind numerous technological activities. New innovations are constantly making discoveries about data, and how we use and analyze it. The article focuses on studying some basic contents about data, big data, and with big data sources which have many outstanding features compared to traditional ones, how we can handle statistical analysis work. Thereby, the article proposes some solutions to promote the application and exploitation of big data sources in statistics in the coming time.

Keywords: Data, big data, statistics.

JEL classification: C55; C8; C82.

1. Đặt vấn đề

Thống kê là khoa học thường nghiên cứu dữ liệu số lớn để đưa ra các nhận xét, đặc điểm, tính chất,... của hiện tượng. Do vậy, dữ liệu lớn (DLL/Big data) xuất hiện cùng với sự phát triển của internet cùng mạng lưới công nghệ thông tin (CNTT) có sự tác động không nhỏ tới công tác thống kê. Mặt khác, thống kê là phương pháp được áp dụng ở hầu hết các lĩnh vực có liên quan đến việc ra quyết định; để có thể đưa ra các kết luận chính xác từ một bộ phận so với dữ liệu toàn bộ và đưa ra quyết định khi đối mặt với kết luận không chắc chắn. Ngày nay, việc sử dụng máy tính hiện đại đã tính toán nhanh các tính toán thống kê quy mô lớn, và cũng đã có những phương

pháp mới tuy nhiên vẫn không thể thay thế được các bước trong quá trình nghiên cứu của thống kê. Trong giai đoạn hiện nay, sự ra đời của DLL chính là cơ hội để thúc đẩy nền kinh tế phát triển và thống kê tiếp tục là một lĩnh vực nghiên cứu thiết thực, vấn đề làm sao để phân tích DLL, thống kê truyền thống cần có điều chỉnh gì cho phù hợp bối cảnh hiện tại, làm sao để khai thác hiệu quả nhất nguồn dữ liệu này, biết cách tận dụng triệt để nguồn dữ liệu này để nắm bắt được cơ hội trong kỷ nguyên số nghiên cứu sẽ góp phần làm rõ một số vấn đề trên.

2. Phương pháp nghiên cứu

Bài viết được tác giả sử dụng kết hợp nhiều phương pháp khác nhau để thu thập, tìm

hiểu, nghiên cứu các tài liệu về DLL và công tác khai thác, ứng dụng DLL trong thực tiễn. Trên cơ sở kế thừa các nghiên cứu về Big data, các phân tích, ứng dụng cơ bản về DLL tác giả thu thập các tài liệu thứ cấp trên các tài liệu đã được công bố từ đó có các sắp xếp, phân tổ, chia nhóm và lại tổng hợp một cách logic, khoa học theo các nhóm chủ đề. Dựa trên các tài liệu đã được thu thập, tổng hợp, tác giả đã áp dụng phương pháp phân tích tư duy logic, phương pháp phân tích chuỗi thời gian kết hợp các chỉ tiêu trong phân tích thống kê mô tả gắn theo các nhóm đối tượng để đạt mục tiêu nghiên cứu đã đề ra.

3. Một số vấn đề cơ bản về dữ liệu lớn

Dữ liệu là gì

Dữ liệu là một tổ hợp các thông tin bao gồm chữ, số, hình ảnh, ... qua đó giúp con người hình dung được tổng thể của sự vật sự việc, nó ứng dụng nhiều trong hầu hết các lĩnh vực diễn ra xung quanh chúng ta, đặc biệt là các ngành kinh tế, công nghệ, kỹ thuật và khoa học.

Dữ liệu đề cập đến các phần thông tin riêng biệt, thường được định dạng và lưu trữ theo cách phù hợp với mục đích cụ thể. Dữ liệu có thể tồn tại ở nhiều dạng khác nhau: dưới dạng số liệu, văn bản được ghi trên giấy, dưới dạng bit hoặc byte được lưu trữ trong bộ nhớ của thiết bị điện tử hoặc dưới dạng dữ kiện trong tâm trí của một người. Tuy nhiên, kể từ khi khoa học máy tính ra đời vào giữa những năm 1900, dữ liệu thường đề cập đến thông tin được truyền hoặc lưu trữ dưới dạng điện tử.

Dữ liệu lớn là gì?

Dữ liệu lớn là thuật ngữ được bắt đầu sử dụng từ giữa những năm 1990 của thế kỷ trước bởi John Mashey, theo đó thuật ngữ này chỉ việc xử lý và phân tích các tập DLL (Diebold, 2012).

Dữ liệu lớn là các tập dữ liệu rất lớn và/hoặc rất phức tạp mà những phương pháp

hiện tại của CNTT chưa phân tích và xử lý tốt được chúng. DLL là các bộ dữ liệu khổng lồ, đòi hỏi kỹ thuật mới để xử lý, lưu trữ và hiểu ý nghĩa của chúng (Batty, 2015). Đó là một khối lượng dữ liệu có cấu trúc và phi cấu trúc quá lớn để xử lý bằng cách sử dụng các công nghệ phần mềm và cơ sở dữ liệu truyền thống; là tập hợp các dữ liệu có quy mô cực kỳ lớn, mang tính phong phú và biến động nhanh.

Tính chất của Big Data - Dữ liệu lớn

Tên gọi dữ liệu lớn khiến đa số chúng ta chỉ hình dung tính chất lớn mà không hình dung về tính phức tạp, nhưng hai tính chất này ở dữ liệu lớn luôn đi cùng nhau, trong đó tính chất “phức tạp” còn đặc trưng và thách thức hơn vấn đề về độ lớn của dữ liệu.

Ban đầu, theo nhà nghiên cứu Laney Doug (Laney Doug, 2001), DLL được đặc trưng bởi 3 tính chất: khối lượng (lượng dữ liệu khổng lồ), vận tốc (dữ liệu được tạo ra trong thời gian thực) và tính đa dạng (dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc). Sau đó, qua nghiên cứu của Mar và McNulty (Marr, 2014) và (McNulty, 2014) đã bổ sung một số tính chất trong đó tập trung về tính chính xác và tính giá trị. Đến nay, DLL thường được đề cập với 5 tính chất (5Vs) cơ bản như sau: *Khối lượng (Volume)*; *Tốc độ (Velocity)*; *Đa dạng (Variety)*; *Độ chính xác (Veracity)*; *Giá trị thông tin (Value)*.

4. Làm thế nào để thống kê, khai thác dữ liệu lớn?

Thống kê nghiên cứu hệ thống các phương pháp thu thập, tổng hợp, phân tích các con số của hiện tượng KTXH số lớn để tìm hiểu bản chất và tính quy luật của hiện tượng trong điều kiện thời gian, địa điểm cụ thể. Như vậy, mục đích của thống kê là dựa trên các thông tin, dữ liệu được thu thập cần xử lý, tổng hợp, trình bày số liệu và tính toán các đặc trưng của đối tượng nghiên cứu nhằm phục vụ cho quá trình phân tích, dự đoán để phục vụ công tác ra quyết định trong quản lý.

Khai thác được DLL - Cơ hội và thách thức với phát triển ngành thống kê ứng dụng:

Dữ liệu trong nghiên cứu thống kê có vai trò rất quan trọng, được coi như nguyên vật liệu đầu vào của quá trình sản xuất thông thường. Hiện nay, nguồn dữ liệu rất lớn, tồn tại hiện hữu xung quanh chúng ta, vậy làm thế nào để khai thác các dữ liệu này hiệu quả là một thách thức không nhỏ với ngành thống kê. Dữ liệu là nguồn chứa hầu hết mọi thông tin của con người, nhưng những thông tin này không lộ ra cho ta dùng ngay mà ta chỉ có thể tìm ra chúng khi phân tích (xử lý) được dữ liệu, nhưng việc xử lý DLL là một vấn đề khó, và hiện nay con người chưa có cách làm được tốt việc này. Tuy vậy, việc ứng dụng và khai thác nguồn DLL vào hoạt động thống kê đang là xu hướng tất yếu trong hoạt động thống kê hiện đại.

Số liệu được thu thập thường rất nhiều và hỗn độn, các dữ liệu đó chưa đáp ứng được cho quá trình nghiên cứu. Để có hình ảnh tổng quát về tổng thể nghiên cứu, số liệu thu thập phải được xử lý tổng hợp, trình bày, tính toán các số đo; kết quả có được sẽ giúp khái quát được đặc trưng của tổng thể sau đó dùng kết quả đó để phân tích, đưa ra các nhận xét, đặc điểm của hiện tượng. Với dữ liệu lớn, quá trình nghiên cứu thống kê đó còn gọi là quá trình phân tích dữ liệu. Vậy phân tích DLL là quá trình thu thập, sắp xếp và tổng hợp các bộ dữ liệu lớn để khám phá các mẫu hoặc thông tin hữu ích khác.

Lợi ích của việc tận dụng được DLL là rất lớn, nhờ vào khả năng khai thác và xử lý lượng dữ liệu khổng lồ, Big Data có thể xác định nhu cầu và dự đoán xu hướng, giúp cho việc đưa ra quyết định đúng đắn và kịp thời. Cụ thể:

Đưa ra quyết định hợp lý dựa trên dữ liệu.

Dữ liệu chính xác, được quản lý tốt dẫn đến việc phân tích chính xác và đưa ra các quyết định hợp lý.

Để duy trì tính cạnh tranh, các doanh nghiệp cần nắm bắt toàn bộ giá trị của DLL. Việc đưa ra quyết định phải dựa trên bằng chứng được trình bày bởi big data chứ không phải bằng cảm tính.

Phân tích dữ liệu là quá trình phát hiện, giải thích và truyền đạt các mô hình có ý nghĩa trong dữ liệu nhằm đưa ra các báo cáo, các bản thống kê tổng hợp theo mong muốn của người dùng. Với sự ra đời của các phương tiện truyền thông xã hội, các tổ chức, doanh nghiệp và người bán hàng đã có nhiều lựa chọn các kênh bán hàng, kênh truyền thông, kênh quảng bá và công cụ marketing mới thay thế cho các phương thức truyền thông. Thông tin dữ liệu thống kê truyền thống thường quan tâm đến yêu cầu tính chính xác, đầy đủ (độ lớn) và thời gian kịp thời (tốc độ) nhưng với DLL thì có những đặc tính, yêu cầu cao hơn với 5 yếu tố nổi bật (5V). Trước đây, nghiên cứu thống kê rất khó khăn khi bộ số liệu đảm bảo tính đủ lớn để có thể nghiên cứu đưa ra những kết luận về hiện tượng nghiên cứu một cách khách quan, chính xác nhất nhưng hiện nay với nguồn DLL thì yêu cầu về tính lớn của bộ dữ liệu đã được đảm bảo, điều này tạo điều kiện rất thuận lợi cho công tác nghiên cứu thống kê. DLL chứa rất nhiều thông tin quý giá, nếu có khả năng khai thác và xử lý lượng dữ liệu khổng lồ đó và biết trích xuất thành công, nó sẽ giúp các nhà nghiên cứu cơ hội tiếp cận và giải quyết rất nhiều vấn đề hóc búa mà trước đây không thể hoặc rất khó thực hiện từ đó sẽ rất hữu ích cho việc kinh doanh, dự đoán các dịch bệnh sắp phát sinh và thậm chí là cả việc xác định điều kiện giao thông, có thể xác định nhu cầu và dự đoán xu hướng, giúp cho việc đưa ra quyết định đúng đắn và kịp thời. Lượng dữ liệu này giúp các đơn vị đạt được tốc độ tăng trưởng nhanh chóng bằng cách phân tích dữ liệu trong thời gian thực. Tuy vậy, cũng chính bởi đặc tính này cũng là

thách thức không nhỏ với công tác thống kê nên rất cần có những điều chỉnh phù hợp trong bối cảnh mới. Bên cạnh đó, các dữ liệu, thông tin được tạo ra trên các phương tiện truyền thông xã hội thường có đặc trưng là không đầy đủ, không hoàn chỉnh, không có cấu trúc rõ ràng hoặc bởi nhiều ngôn ngữ trộn lẫn cũng theo đó xuất hiện nhiều, đòi hỏi các công cụ thống kê phân tích dữ liệu cũng có sự thay đổi và cải tiến để theo kịp các yêu cầu của người dùng.

Quy trình nghiên cứu thống kê dữ liệu lớn

Phân tích DLL là quá trình thu thập, sắp xếp và tổng hợp các bộ dữ liệu lớn để khám phá các mẫu hoặc thông tin hữu ích khác.

Trong nghiên cứu thống kê dữ liệu lớn cũng có 2 hướng nghiên cứu:

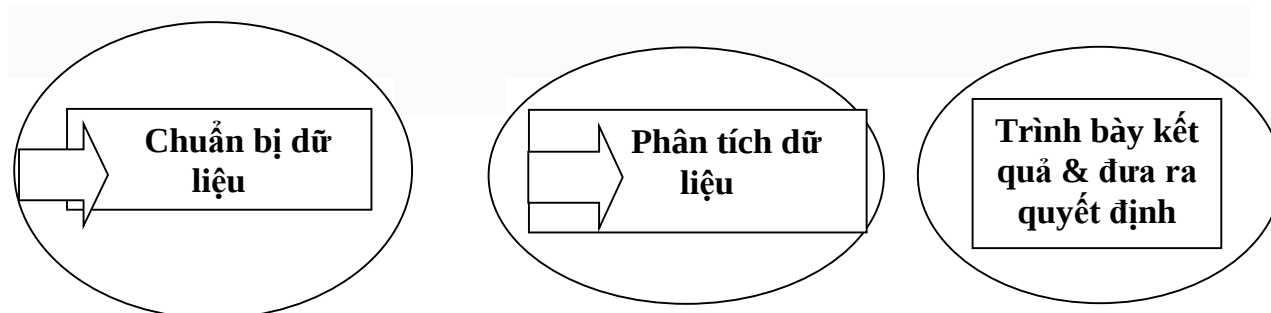
Phân tích dữ liệu trong nghiên cứu định tính.

Phân tích và nghiên cứu dữ liệu thông tin định tính bao gồm các thông tin từ ngữ, sự mô tả, hình ảnh, đồ vật... nên việc thu thập

thông tin này khá khó khăn, do đó, phương pháp này thường được sử dụng để nghiên cứu khám phá cũng như phân tích dữ liệu. Quan trọng trong nghiên cứu định tính là tìm kiếm các mẫu trong dữ liệu định tính. Đặc biệt, quy trình phân tích dữ liệu trong nghiên cứu định tính được thực hiện thủ công. Ở đây, các chuyên gia đọc thông tin có thể truy cập và tìm các từ đơn điệu.

Phân tích dữ liệu trong nghiên cứu định lượng

Phân tích dữ liệu định lượng được sử dụng với các thông tin được lượng hóa và phương pháp này thường được phổ biến hơn. Một quy trình phân tích dữ liệu thông thường có 3 giai đoạn gồm: Tìm hiểu thu thập dữ liệu; Phân tích dữ liệu và phân loại dữ liệu; Cuối cùng là tạo các báo cáo và các quyết định quản lý dựa trên dữ liệu phân tích. Có thể mô tả bài toán phân tích dữ liệu một cách tổng quát như sau:



Như vậy, để khai thác được hiệu quả nguồn dữ liệu khổng lồ trên, quá trình khai thác cũng cơ bản tuân theo các giai đoạn, tuy vậy, nội dung cụ thể của từng bước có thể khác nhau phụ thuộc vào nguồn dữ liệu cũng như mục đích nghiên cứu.

Giai đoạn 1: Chuẩn bị dữ liệu để phân tích.

Xây dựng chiến lược Big Data -> Xác thực dữ liệu -> Xác định nguồn Big Data -> Truy cập, quản lý và lưu trữ Big Data.

Nguồn: Tác giả
Việc chuẩn bị dữ liệu bao gồm những bước sau đây.

Xây dựng chiến lược Big Data: Ở cấp độ cao, trước tiên cần xây dựng chiến lược tạo nguồn big data từ đó xây dựng kế hoạch được thiết kế để giúp giám sát và cải thiện cách thu thập, lưu trữ, quản lý, chia sẻ và sử dụng dữ liệu cho đơn vị sử dụng.

Xác thực dữ liệu

Đối với nghiên cứu thống kê định lượng, việc phân tích mô tả thường đưa ra những con số tối ưu. Tuy nhiên, phân tích không bao giờ

đủ để chỉ ra lý do ẩn sau những con số này. Điều quan trọng là phải chọn ra kỹ thuật nào sẽ được sử dụng để nghiên cứu và phân tích dữ liệu phù hợp với khảo sát đánh giá, do vậy, với dữ liệu bên cạnh việc xác thực cần chỉnh sửa dữ liệu và mã hóa dữ liệu.

Xác định nguồn DLL: Về xác định các nguồn Big Data, trước hết cần xác định rõ các loại dữ liệu:

Siêu dữ liệu (Metadata).

Dữ liệu thô (Raw data).

Dữ liệu truyền trực tiếp: dữ liệu từ Internet vạn vật (IoT) và các thiết bị được kết nối truyền vào hệ thống công nghệ thông tin từ thiết bị như điện thoại thông minh, ô tô thông minh.

Truyền thông xã hội (Facebook, YouTube, Instagram, ...): DLL ở dạng hình ảnh, video, giọng nói, văn bản và âm thanh, rất hữu ích cho chức năng tiếp thị, bán hàng và hỗ trợ. Dữ liệu này thường ở dạng phi cấu trúc hoặc bán cấu trúc, vì vậy nó đặt ra một thách thức riêng cho việc tiêu thụ và phân tích.

Dữ liệu có sẵn được công bố.

Dữ liệu khác: nguồn dữ liệu khác đến từ khách hàng, nhà cung cấp hoặc dữ liệu đám mây.

Truy cập, quản lý và lưu trữ dữ liệu lớn

Các hệ thống máy tính hiện đại cung cấp tốc độ, sức mạnh và tính linh hoạt cần thiết để nhanh chóng truy cập số lượng lớn và các loại DLL. Một số dữ liệu lớn có thể được lưu trữ tại chỗ trong kho dữ liệu truyền thống nhưng cũng có các tùy chọn linh hoạt, chi phí thấp để lưu trữ và xử lý dữ liệu lớn thông qua các giải pháp đám mây, hồ dữ liệu, đường ống dữ liệu và Hadoop. Trong giai đoạn này sử dụng phương pháp khai phá dữ liệu (Data mining), đây là quá trình phân loại, sắp xếp các tập hợp dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu.

Bước 1: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có tạp âm hay bất thường trong dữ liệu.

Bước 2: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu, nhiều nguồn dữ liệu sẽ kết hợp lại thành một.

Bước 3: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.

Bước 4: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.

Giai đoạn 2: Tiến hành phân tích dữ liệu.

Với các công nghệ hiệu suất cao như điện toán lưới hoặc phân tích trong bộ nhớ, các đơn vị có thể chọn sử dụng tất cả DLL của họ để phân tích. Ngày nay, nguồn cấp DLL cung cấp các công cụ phân tích nâng cao như trí tuệ nhân tạo (AI) và máy học. Trong bước này, tiếp tục sử dụng phương pháp khai phá dữ liệu để trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có. Hơn nữa, khai phá dữ liệu không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.

Giai đoạn 3: Trình bày kết quả và đưa ra quyết định.

Trình bày thông tin – Trong giai đoạn cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận.

5. Nghiên cứu ứng dụng thống kê dữ liệu lớn trong thống kê chỉ số giá

Giá tiêu dùng và chỉ số giá tiêu dùng đóng vai trò rất quan trọng trong công tác quản lý và điều hành các chính sách vĩ mô của Nhà nước, góp phần phát triển hoạt động sản xuất kinh doanh và trao đổi thương mại quốc tế. Tại các quốc gia phát triển trên thế giới hiện nay đã ứng dụng khai thác dữ liệu lớn (big data) để tính chỉ số giá tiêu dùng và đạt được những hiệu quả nổi trội. Ở Việt Nam, ứng dụng này cũng mới được thực hiện ở Thành

phố Hồ Chí Minh và tiến tới nhân rộng ứng dụng này trong phạm vi cả nước.

Thống kê chỉ số giá tiêu dùng dựa trên số liệu thu thập trực tiếp.

Giá tiêu dùng và chỉ số giá tiêu dùng đóng vai trò rất quan trọng trong công tác quản lý và điều hành các chính sách vĩ mô của Nhà nước góp phần phát triển hoạt động sản xuất kinh doanh và trao đổi thương mại quốc tế.

Chỉ số giá tiêu dùng là chỉ tiêu tương đối phản ánh xu hướng và mức độ biến động giá của "rổ" hàng hóa và dịch vụ tiêu dùng đại diện nói trên, khi giá của các mặt hàng, nhóm hàng trong "rổ" có thay đổi. Chỉ số này so sánh mức độ biến động giá của các mặt hàng đại diện trong kỳ báo cáo so với kỳ gốc. Giá của rổ hàng hóa của kỳ gốc được quy định là 100 và giá của các kỳ khác được biểu hiện bằng tỷ lệ phần trăm so với giá kỳ gốc.

Để tính Chỉ số giá tiêu dùng/tháng cần thực hiện các bước sau đây:

Lập bảng giá kỳ gốc.

Lập bảng quyền số cố định kỳ gốc.

Thu thập giá bán lẻ của các mặt hàng và dịch vụ đại diện

Tính giá bình quân hàng tháng theo từng khu vực của các tỉnh thành phố.

Tính chỉ số giá cấp tỉnh/thành phố theo từng khu vực và chung cả tỉnh.

Tính chỉ số giá cấp vùng kinh tế theo từng khu vực và chung cả vùng.

Tính chỉ số giá cả nước theo từng khu vực thành thị, nông thôn và chung cả nước.

Để có chỉ số giá tiêu dùng phản ánh đúng mức độ biến động của giá cả trên thị trường, việc thu thập giá đóng vai trò rất quan trọng. Chính vì vậy, từ nhiều năm nay, phương pháp điều tra giá đã được xác định là điều tra trực tiếp, do các điều tra viên ở các tỉnh/thành phố thực hiện. Tại Việt Nam, thông tin thống kê về chỉ số giá tiêu dùng được thu thập từ cuộc điều tra giá tiêu dùng do Tổng cục Thống

kê triển khai ở cả 63 tỉnh, thành phố và được công bố hằng tháng vào các ngày cuối tháng. Trong phương án điều tra giá tiêu dùng giai đoạn 2014-2019, ngành Thống kê thực hiện cuộc điều tra giá tiêu dùng bằng phương pháp chọn mẫu đối với rổ hàng hóa, dịch vụ gồm 654 mặt hàng tại các sạp, quầy hàng tại các chợ, điểm bán hàng (chuyên bán lẻ), các cơ sở kinh doanh dịch vụ, ... có địa điểm kinh doanh ổn định. Hàng hóa và dịch vụ trong rổ hàng hóa được chia làm 3 nhóm chính: Nhóm thứ nhất chỉ điều tra 1 lần trong tháng và sẽ điều tra vào ngày 10 hàng tháng; nhóm thứ hai sẽ điều tra 3 lần trong tháng vào các ngày 1, 10, 20 hàng tháng; nhóm thứ ba theo số lần phát sinh trong tháng.

Trong những năm gần đây, việc tính chỉ số đã được thực hiện bằng chương trình phần mềm máy tính chuyên dụng nên chỉ số giá tiêu dùng ngày càng có độ tin cậy cao, nhưng với phương pháp tính dựa trên việc thu thập số liệu trực tiếp, phương pháp điều tra truyền thống cũng còn một số bất cập.

Thứ nhất, công tác thu thập tại địa bàn gặp nhiều khó khăn, đặc biệt khi thời gian thu thập đúng vào những ngày nghỉ Lễ, Tết, thời gian địa phương thực hiện giãn cách xã hội... khiến phần lớn các cơ sở kinh doanh không mở cửa bán hàng và giá cả hàng hóa, dịch vụ thường có sự biến động lớn tại những thời điểm này cũng là một hạn chế cần phải xử lý.

Thứ hai, vẫn còn phát sinh sai số phi chọn mẫu trong quá trình triển khai thu thập.

Thứ ba, khó khăn trong việc xử lý đối với những hàng hóa và dịch vụ có chu kỳ sống ngắn hạn, không tồn tại vào thời điểm điều tra và nhiều hàng hóa mới phát sinh trong kỳ điều tra.

Ngoài ra, những vấn đề về kinh phí điều tra, việc chọn mẫu và tính toán số lượng các điểm bán hàng bình ổn ở địa phương, việc xử lý các cửa hàng trong mẫu điều tra ngày càng thu hẹp kinh doanh, thị phần giảm sút ... cũng

là những khó khăn khi triển khai thu thập số liệu theo phương pháp điều tra truyền thống.

Thống kê chỉ số giá tiêu dùng từ khai thác nguồn dữ liệu lớn.

Nền tảng của cuộc cách mạng công nghiệp 4.0, cùng với sự phát triển của kinh tế số đã tạo ra những nguồn dữ liệu mới – đây là cơ hội tuyệt vời cho ngành thống kê cải thiện chất lượng, nâng cao hiệu quả công tác thu thập thông tin. Nguồn dữ liệu giá thu thập trực tuyến này sẽ giúp công tác thống kê giá đo lường chính xác hơn sự thay đổi giá, mở rộng cỡ mẫu, xác định chính xác các mặt hàng thay thế theo hành vi người tiêu dùng, đo lường chính xác sự thay đổi giá, giảm hoặc loại bỏ sự phụ thuộc từ người trả lời và giảm chi phí điều tra. Ngoài ra, nguồn dữ liệu thu thập từ các trang Web sẽ giúp cải thiện thời gian thu thập, dữ liệu chi tiết hơn, nhiều hơn và đa dạng hơn, tần suất thu thập dữ liệu nhiều hơn mà không gia tăng chi phí.

Với ứng dụng DLL để thống kê chỉ số giá, đã tiến hành nghiên cứu thử nghiệm, thu thập thông tin từ 28 trang web của các công ty và các sàn thương mại điện tử với 246.069 số lượng mặt hàng. Nguồn dữ liệu giá thu thập trực tuyến sẽ mang lại cơ hội tốt cho ngành thống kê xử lý các thách thức mà thống kê giá tiêu dùng truyền thống đang đối mặt. Phương pháp thu thập giá từ DLL có số lượng mặt hàng lớn hơn rất nhiều so với phương pháp truyền thống, khoảng 250 ngàn mặt hàng vào 3 kỳ/tháng nên chỉ số CPI được tính toán từ dữ liệu lớn thể hiện biến động giá của thị trường nhạy hơn và ổn định hơn. Thực tế, nếu một mặt hàng trong rổ hàng hóa của CPI truyền thống (đặc biệt là mặt hàng chỉ thu thập 1 kỳ/1 tháng) biến động mạnh (tăng cao hoặc giảm sâu) thì sẽ ảnh hưởng rất lớn đến CPI chung. Trong khi đó, đối với DLL sẽ không ảnh hưởng nhiều vì có rất nhiều hàng hóa khác cùng nhóm được thu thập nhưng không có sự biến động lớn.

Ngoài ra, phương pháp thu thập từ DLL có tính đại diện mẫu cao hơn vì thu thập giá của tất cả các mặt hàng, do đó có thể xem mẫu gần như là đại diện cho toàn bộ các hàng hóa có trên thị trường. Ngược lại, đối với CPI truyền thống, rổ hàng hóa được cập nhật theo chu kỳ 5 năm nên có rất nhiều mặt hàng tiêu dùng phổ biến xuất hiện trên thị trường nhưng chưa được cập nhật, hơn nữa rổ hàng hóa cũng chỉ có 654 mặt hàng đại diện tiêu dùng phổ biến nên cũng thiếu rất nhiều so với thực tế biến động liên tục của thị trường.

Để tính toán CPI từ nguồn dữ liệu lớn, nghiên cứu đã sử dụng phương pháp nghiên cứu định lượng: Phương pháp thu thập dữ liệu, Phương pháp tổng hợp CPI dựa trên dữ liệu lớn. Dữ liệu được thu thập 3 kỳ trong tháng: Kỳ 1 thu thập vào ngày 01 tháng báo cáo; Kỳ 2 thu thập vào ngày 10 tháng báo cáo; Kỳ 3 thu thập vào ngày 20 tháng báo cáo. Thời gian thu thập dữ liệu: Bắt đầu từ năm 2017.

Quy trình thu thập và xử lý

Bước 1: Vào các ngày 1, 10 và 20 hàng tháng tiến hành thu thập thông tin tại các trang. Các thông tin này đều là dữ liệu phi cấu trúc.

Bước 2: Xử lý dữ liệu từ phi cấu trúc thành có cấu trúc.

Bước này chủ yếu là quá trình “dạy” máy hiểu được ngôn ngữ, bao gồm các nội dung:

(1) Tuyển chọn kỹ thuật viên để xử lý thông tin đã được thu thập vào máy theo cấu trúc quy định; (2) Thiết kế và viết chương trình nhập tin; (3) Tập huấn kỹ thuật viên; (4) Tiến hành cho kỹ thuật viên xử lý thông tin (10% mẫu); (5) Viết chương trình xử lý ngôn ngữ để máy có thể hiểu và xử lý dữ liệu; (6) Kiểm tra dữ liệu sau khi máy đã xử lý; (7) Nếu dữ liệu đạt chất lượng từ 80% trở lên, cho máy xử lý toàn bộ dữ liệu; (8) Nếu chất lượng dữ liệu chưa đạt tới 80%, tiến hành mở thêm 5% mẫu xử lý và quay lại bước tiến hành cho kỹ thuật viên xử lý thông tin.

Bước 3: Xử lý dữ liệu. Kiểm tra giá các mặt hàng biến động (cao hoặc thấp nhiều) so với tháng trước. Gán giá cho các mặt hàng có giá ở kỳ trước nhưng không có giá ở kỳ này (mặt hàng nào gán giá 3 lần thì sẽ bị loại bỏ, không đưa vào tính toán CPI).

Bước 4: Tính CPI

Tính CPI theo các phương pháp và thu được kết quả như sau:

Phương pháp 1: Sử dụng công thức bình quân cộng để tính giá bình quân từ dữ liệu lớn và dùng giá này để tính CPI.

Phương pháp 2: Sử dụng công thức bình quân nhân để tính giá bình quân từ dữ liệu lớn và dùng giá này để tính CPI.

Phương pháp 3: Kết hợp giữa chỉ số giá tính theo phương pháp truyền thống và phương pháp thu thập từ dữ liệu lớn. Phương pháp này tính toán quyền số dựa vào phía cầu

hàng hóa (người tiêu dùng), thu thập từ dữ liệu đại diện cho chi tiêu của người dân ở khu vực thành thị, kết hợp với kết quả tính toán CPI được thu thập từ phương pháp truyền thống ở khu vực nông thôn (đại diện cho chi tiêu của người dân ở khu vực nông thôn). Quyền số kết hợp là tỷ trọng chi tiêu của khu vực thành thị và tỷ trọng chi tiêu của khu vực nông thôn (Theo Tổng cục Thống kê quyền số này là cố định, 5 năm mới thay đổi 1 lần).

Phương pháp 4: Kết hợp giữa chỉ số giá tính theo phương pháp truyền thống và phương pháp thu thập từ dữ liệu lớn. Phương pháp này tính toán quyền số dựa vào phía cung hàng hóa (doanh nghiệp và hộ cá thể bán lẻ hàng hóa và dịch vụ), do đó sẽ tính toán riêng chỉ số giá tiêu dùng do khối doanh nghiệp cung cấp và chỉ số giá tiêu dùng do khối cá thể cung cấp.

Bảng 01: Chỉ số giá tiêu dùng các tháng so tháng trước

(DVT: %)

Quan sát	PP truyền thống	P/A 1	P/A 2	P/A 3	P/A 4
1	100.63	100.66	100.59	100.65	100.65
2	100.17	100.22	100.09	100.21	100.21
3	100.05	100.12	100.11	100.11	100.10
4	100.19	100.49	100.41	100.46	100.42
5	100.34	100.37	100.39	100.36	100.36
6	99.70	99.96	99.97	99.94	99.92
7	100.12	100.01	100.04	100.02	100.04
8	100.43	100.22	100.23	100.27	100.32
9	100.55	100.33	100.33	100.34	100.36
10	99.91	99.89	99.88	99.90	99.92
11	100.48	100.20	100.19	100.21	100.22
12	100.81	100.66	100.68	100.68	100.71
13	100.64	100.44	100.45	100.46	100.49
14	99.59	99.78	99.84	99.76	99.73

Nguồn: Trích từ nghiên cứu thí điểm tại thành phố Hồ Chí Minh

Dữ liệu thu thập từ phương pháp truyền thống đại diện cho khối cá thể và thu thập từ dữ liệu lớn ở phương pháp 1. Quyền số kết hợp là tỷ trọng tổng mức bán lẻ của khối doanh nghiệp và khối cá thể chia chi tiết theo từng nhóm hàng (quyền số này thay đổi hàng năm, dựa vào kết quả điều tra doanh nghiệp hoặc có thể thay đổi theo từng tháng, dựa vào

báo cáo tháng của doanh nghiệp mẫu và hộ cá thể mẫu).

Với kết quả bảng 01, kết quả tính toán CPI từ dữ liệu lớn ở cả 4 phương pháp đều thể hiện đúng xu hướng và không có chênh lệch nhiều so với CPI truyền thống. Trong giai đoạn nghiên cứu, CPI so tháng trước thì có 6 tháng CPI tính theo phương pháp

truyền thống cao hơn CPI tính từ dữ liệu lớn và có 8 tháng thấp hơn. Đặc biệt có 3 tháng, mức chênh lệch này chỉ có 0,02% tháng có mức chênh lệch cao nhất là 0,25%.

Để thấy rõ hơn đặc điểm của các dãy số tương ứng với chỉ số giá cả được thực hiện theo các phương án đề xuất, tác giả nghiên cứu thông qua các chỉ tiêu của thống kê mô tả, qua đó, một không xuất hiện với các trường hợp này, ngoài ra hầu hết các chỉ tiêu

khác từ giá trị bình quân, trung vị giá trị nhỏ nhất và giá trị lớn nhất của phương án 4 có giá trị sát nhất với các giá trị tương ứng của phương án đối sánh, phương pháp thu thập truyền thống. Tuy có chỉ tiêu, phương sai và độ lệch chuẩn thì có độ lệch tương đối giữa 4 phương án. Và để thấy rõ hơn, tác giả nghiên cứu độ chênh lệch giữa các phương án qua các chỉ tiêu qua bảng sau:

Bảng 02: Chênh lệch kết quả của 4 phương án tính CPI so với phương án truyền thống

(DVT: %)

Chỉ tiêu	PP truyền thống	P/A 1	P/A 2	P/A 3	P/A 4
Mean	100.25785	-0.01857	-0.02929	-0.01714	-0.01143
Standard Error	0.097022	-0.0243	-0.02819	-0.02339	-0.02127
Median	100.265	-0.045	-0.055	-0.025	0.005
Standard Deviation	0.3630252	-0.09093	-0.10547	-0.0875	-0.0796
Sample Variance	0.1317873	-0.05775	-0.06545	-0.05587	-0.05145
Kurtosis	-0.6219723	-0.15981	-0.2508	-0.07734	0.073358
Skewness	-0.3919638	0.394866	0.559459	0.332952	0.270007
Range	1.22	-0.34	-0.38	-0.3	-0.24
Minimum	99.59	0.19	0.25	0.17	0.14
Maximum	100.81	-0.15	-0.13	-0.13	-0.1
Sum	1403.61	-0.26	-0.41	-0.24	-0.16
Count	14	0	0	0	0
Confidence Level(95.0%)	0.2096045	-0.0525	-0.06089	-0.05052	-0.04596

Như vậy, với tất cả các chỉ tiêu nghiên cứu, so với CPI truyền thống, CPI được tính theo phương pháp 4 có mức độ chênh lệch thấp nhất trong 4 phương pháp tính và là phương pháp kế thừa, kết hợp giữa phương pháp 1 và CPI tính theo phương pháp truyền thống. Và quan trọng hơn cả là trong tất cả các tháng của giai đoạn phân tích, CPI luôn phản ánh cũng xu hướng tăng hoặc giảm, kết quả tính CPI tương đối giống nhau.

6. Hàm ý chính sách

Mục tiêu ứng dụng DLL để hiện đại hóa, giảm chi phí, nâng cao chất lượng và tăng cường khả năng dự báo đối với một số chỉ tiêu thống kê. Tại Việt Nam, ngày 10 tháng 5 năm 2018, Thủ tướng Chính phủ đã

Nguồn: Theo tính toán và phân tích của tác giả
ban hành Quyết định số 501/QĐ-TTg phê duyệt Đề án ứng dụng công nghệ thông tin - truyền thông trong Hệ thống thống kê nhà nước giai đoạn 2017-2025, tầm nhìn đến năm 2030, trong đó có mục tiêu là ứng dụng DLL vào năm 2020 để hiện đại hóa, giảm chi phí, nâng cao chất lượng và tăng cường khả năng dự báo đối với một số chỉ tiêu thống kê trong các lĩnh vực thống kê giá và thị trường bất động sản và đến năm 2025, sẽ mở rộng ứng dụng DLL trong các lĩnh vực thống kê di cư nội địa, du lịch, vận tải và một số lĩnh vực thống kê khác; đồng thời cũng đề ra nhiều giải pháp để ứng dụng DLL trong hoạt động thống kê nhà nước.

7. Một số kiến nghị góp phần nâng cao khả năng khai thác và ứng dụng dữ liệu lớn trong thời gian tới.

Thứ nhất, đào tạo nguồn nhân lực đảm bảo cả về số lượng và chất lượng để đáp ứng năng lực khai thác nguồn DLL: Khai thác DLL phục vụ công tác thống kê là một nội dung mới, nguồn nhân lực trong lĩnh vực này còn thiếu về số lượng và cũng có những hạn chế nhất định về chất lượng. Công việc vận hành hệ thống, phát triển phần mềm, khai phá dữ liệu... là các kỹ năng cần thiết để thực hiện tốt công việc này. Do đó, các cơ sở giáo dục cần nâng cao chất lượng công tác đào tạo, chương trình đào tạo phải bám sát thực tiễn, đáp ứng được nhu cầu công việc. Trên cơ sở đó giúp người học có thể xử lý được các bài toán thực tiễn về DLL khi được tuyển dụng. Đồng thời, Nhà nước cần xây dựng cơ chế phối hợp, tạo điều kiện thuận lợi để tạo mối quan hệ mật thiết giữa các đơn vị sử dụng lao động với Nhà trường. Bên cạnh đó cần chú trọng đầu tư vào công tác đào tạo nguồn nhân lực chất lượng cao trong lĩnh vực công nghệ thông tin, phân tích dữ liệu và tập trung đào tạo, nâng cao năng lực, kỹ năng, nâng cao trình độ ứng dụng công nghệ thông tin, các kiến thức, kỹ năng có liên quan đến dữ liệu lớn cho cán bộ làm công tác nghiệp vụ để có thể nghiên cứu ứng dụng khoa học dữ liệu và các công nghệ tiên tiến hiện đại, áp dụng phù hợp cho công tác thống kê tại Việt Nam.

Thứ hai, xây dựng cơ sở hạ tầng CNTT phát triển mạnh và các công nghệ nổi trội.

Việc khai thác, phân tích dữ liệu lớn đòi hỏi phải có cơ sở hạ tầng CNTT phát triển mạnh và các công nghệ nổi trội. Muốn phát triển lĩnh vực này cần phải tập trung đầu tư phát triển về hạ tầng tính toán: nâng cấp hạ tầng CNTT, tiến hành thuê hạ tầng CNTT phục vụ hệ thống thu thập thông tin bằng phiếu điện tử; nâng cao năng lực Trung tâm máy chủ; chú trọng triển khai hệ thống thu

thập thông tin trực tuyến trên các trang web và xử lý thông tin theo thời gian thực; xây dựng kiến trúc tổng thể để hệ thống có độ mở, sẵn sàng tích hợp và nâng cấp mở rộng hệ thống khi cần thiết. Hỗ trợ, tạo động lực để các DN, các đơn vị nghiên cứu, cung cấp các giải pháp phần mềm và công cụ hỗ trợ trong việc lưu trữ, quản lý, tích hợp và phân tích dữ liệu. Chẳng hạn như việc thiết lập và xây dựng Hệ sinh thái Hadoop (Hadoop được coi là xương sống trong các dự án DLL).

Ngoài ra, hiện tại trong Luật Thống kê mới chỉ quy định 3 nguồn dữ liệu chính thức phục vụ sản xuất thông tin thống kê trong hệ thống thống kê nhà nước đó, đó là: Điều tra thống kê, dữ liệu hành chính và chế độ báo cáo thống kê. Do vậy, để DLL có thể chính thức được áp dụng với tư cách là một nguồn thông tin chính thống trong sản xuất số liệu thống kê trong hệ thống thống kê nhà nước thì cần sớm hoàn thiện khuôn khổ pháp lý liên quan đến DLL, cần điều chỉnh Luật để có thể coi DLL là một trong các nguồn dữ liệu chính thức trong hệ thống thống kê nhà nước.

8. Kết luận

Dữ liệu lớn đánh dấu một bước quan trọng trong việc tìm kiếm của con người để định lượng và hiểu thế giới, một ưu thế của những thứ chưa bao giờ được đo lường, lưu trữ, phân tích và chia sẻ trước khi được dữ liệu hóa. DLL đang thay đổi thế giới, nó đã hoàn toàn biến chuyển cách chúng ta sống, cách chúng ta làm và cách chúng ta nghĩ... DLL sẽ giúp chúng ta đưa ra hành động hợp lý, tối ưu hóa được nhiều vấn đề, quan trọng là biết cách khai thác, ứng dụng trong thống kê để đưa ra được các quyết định hợp lý. Tóm lại, cùng với sự phát triển của kỹ nguyên số, sự ra đời của DLL chính là cơ hội thống kê phát triển qua đó góp phần thúc đẩy nền kinh tế phát triển nếu chúng ta biết cách tận dụng triệt để.

TÀI LIỆU THAM KHẢO

- [1]. Bảo Lam. (2018). *Công nghiệp 4.0 và thách thức Big Data* <https://diendandoanhnghiep.vn/cong-nghiep-4-0-va-thach-thuc-big-data-128866.html>. Truy cập ngày 20 tháng 12 năm 2022.
- [2]. Batty, M. (2015a). *Does big data lead to smarter cities? Problems, pitfalls and opportunities*. I/S, A Journal of Law and Policy for the Information Society, 11(1), pp. 127–151.
- [2]. Diebold F. (2012). A Personal Perspective on the Origin(s) and Development of “Big Data”: The Phenomenon, the Term, and the Discipline, Second Version," PIER Working Paper Archive 13-003, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, revised 26 Nov 2012.
- [3]. Gil Allouche. (2014). <https://www.emergingedtech.com/2014/06/how-big-data-can-help-improve-teaching/> Truy cập ngày 18 tháng 12 năm 2022.
- [4]. Laney Doug. (2001). *3-D Data Management: Controlling Data Volume, Velocity and Variety*, Magazine Article, Application Delivery Strategies by META Group Inc, Vo 949, Feb 6 2001.
- [5]. Marr B. (2014). *Big data: The 5 vs everyone must know*. March 6. Available at: <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>
- [6]. McNulty E. (2014). *Understanding Big Data: The seven V's*. 22 May. Available at: <http://dataconomy.com/seven-vs-big-data/> (accessed 4 September 2015).
- [7]. Quyết định số 501/QĐ-TTg ngày 10 tháng 5 năm 2018 của Thủ tướng Chính phủ về việc phê duyệt: *Đề án ứng dụng công nghệ thông tin - truyền thông trong Hệ thống thống kê Nhà nước giai đoạn 2017 - 2025, tầm nhìn đến năm 2030*.
- [8]. Hà Văn Sơn, Nguyễn Thanh Bình. (2021). *Chỉ số giá tiêu dùng từ góc nhìn dữ liệu lớn*. Tạp chí Diễn đàn Doanh nghiệp số tháng 12 năm 2021.
- [9]. *Tổng quan về Big Data và những vấn đề liên quan* <https://techinsight.com.vn/tong-quan-ve-big-data-va-nhung-van-de-lien-quan> Truy cập ngày 18 tháng 12 năm 2022.
- [10]. Trần Thị Thu. (2019). Vai trò của các cơ quan Thống kê đối với Big Data. *Tạp chí Con số và sự kiện*, số tháng 6 năm 2019.
- [11]. Trịnh Thu Trang. (2019). Nghiên cứu về lợi ích của dữ liệu lớn – Big Data với doanh nghiệp thương mại điện tử trong nước và thế giới. *Tạp chí Công thương số tháng 8 năm 2019*.

Thông tin tác giả:

Nguyễn Thị Nhung

- Đơn vị công tác: Khoa Kinh tế, Trường Đại học Kiến trúc Đà Nẵng
- Địa chỉ email: nhungnt@tueba.edu.vn

Ngày nhận bài: 31/3/2023

Ngày nhận bản sửa: 8/5/2023

Ngày duyệt đăng: 25/5/2023