ỨNG DỤNG BIG DATA TRONG THỐNG KÊ GIÁ

CN.Nguyễn Thị Minh Ánh
Phòng Nghiên cứu khoa học và Chiến lược PTTK, Viện KHTK

Dữ liệu lớn (Big data) là chủ đề đang thu hút được sự quan tâm tại nhiều quốc gia trên thế giới với những lĩnh vực ứng dụng cụ thể như: chính trị, kinh tế, giao thông vận tải, y học, thống kê… Để hiểu rõ hơn về những ứng dụng Big data vào công tác thống kê của các cơ quan và tổ chức thống kê thế giới, bài viết sau giới thiệu về hai dự án thực tế khai thác loại dữ liệu này trong lĩnh vực thống kê giá mà Cơ quan thống kê quốc gia Anh (ONS) và Viện Thống kê và nghiên cứu Kinh tế quốc gia Pháp (INSEE) đã thực hiện thành công.

1. Sử dụng công nghệ Web Scraper để khai thác nguồn dữ liệu Big data phục vụ tính chỉ số CPI

Sự tăng trưởng của bán hàng trực tuyến những năm gần đây đồng nghĩa với việc các dịch vụ hàng hóa và các thông tin giá cả liên quan có thể được tìm thấy thông qua mạng. Thông tin chỉ số giá tiêu dùng CPI và chỉ số giá bán lẻ RPI là các chỉ tiêu kinh tế quan trọng mà ONS đặc biệt quan tâm. Với sự hỗ trợ của công nghệ Web Scraper (công cụ phần mềm giúp trích xuất dữ liệu từ các trang web) cùng với kỹ thuật trích xuất nội dung từ trang web (Web scraping) có thể mang lại cơ hội cho ONS thu thập các dữ liệu về một số mặt hàng và dịch vụ một cách tự động thay vì việc các điều tra viên phải đi điều tra từng cửa hàng để thu thập số liệu. Điều này mang lại một loạt các lợi ích tiềm năng như giảm bớt chi phí thu thập thông tin, mở rộng phạm vi (ví dụ: nhiều danh mục hàng hóa sản phẩm hơn) và tăng tính đều đặn.

Cơ quan thống kê quốc gia Anh (ONS) đã tiến hành 4 dự án về Big data cho thống kê Nhà nước, trong đó có dự án "Sử dụng công nghệ Web Scraper để khai thác nguồn dữ liệu Big data phục vụ tính chỉ số CPI". Dự án kéo dài trong vòng 15 tháng và đã kết thúc vào tháng 3/2015.

Mục tiêu của dự án nhằm khai thác nguồn dữ liệu Big data thông qua công nghệ Web Scraper để ứng dụng vào lĩnh vực thống kê giá. Đồng thời phát triển các phương pháp phân tích, xử lý nguồn dữ liệu Big data thu thập được. Trên cơ sở kết quả đó sẽ được so sánh với những kết quả có được từ việc thu thập dữ liệu theo phương pháp truyền thống, giúp cho người khai thác thấy được những ưu điểm, hạn chế cũng như tính toán được chi phí, lợi ích đối với việc ứng dụng nguồn Big data trong thống kê Nhà nước.

Quy trình thực hiện dự án được tiến hành theo 3 bước:

Bước 1: Xác định nguồn dữ liệu thu thập

Dữ liệu cần thu thập là dữ liệu liên quan đến giá như: mức giá, lượng bán, doanh thu mặt hàng… của các mặt hàng tiêu dùng. Nguồn dữ liệu được lấy từ hệ thống thông tin của các siêu thị hàng hóa bán lẻ trực tuyến.

Việc lựa chọn siêu thị cụ thể và số lượng siêu thị được dựa trên việc xem xét các tiêu chí như: doanh thu, thị phần của siêu thị trên thị trường, các mặt hàng mà siêu thị cung cấp, khả năng cung cấp thông tin của siêu thị, kinh phí thực hiện dự án...

Số lượng cũng như các mặt hàng tiêu dùng mà dự án thu thập được căn cứ theo rổ hàng hóa phục vụ cho việc tính CPI.

Trên thực tế, dự án "Sử dụng công nghệ Web Scraper để khai thác nguồn dữ liệu Big data phục vụ tính chỉ số CPI" của ONS đã tiến hành thu thập các thông tin liên quan đến giá cả của 35 mặt hàng tiêu dùng thuộc rổ hàng hóa CPI của 3 siêu thị bán hàng trực tuyến.

Bước 2: Lắp đặt, vận hành công cụ Web Scraper phục vụ việc thu thập dữ liệu

Các thông tin liên quan tới giá sản phẩm tiêu dùng được có được thông qua việc truy cập vào các kho dữ liệu của 3 siêu thị tiến hành thử nghiệm. Tuy nhiên, thực tế những dữ liệu này chỉ là những dữ liệu thô gồm cả dữ liệu có cấu trúc và phi cấu trúc (như các bản báo giá, đơn đặt hàng, phiếu thanh toán, hóa đơn… gồm cả dữ liệu hình ảnh, số liệu, kí tự…). Đây là những thông tin cần thiết phục vụ cho việc tính CPI (như giá cả, hay lượng bán đều nằm trong những dữ liệu thô này). Vì vậy để có được những thông tin đáp ứng được đúng nhu cầu của người dùng tin thì cần phải lắp đặt một công cụ có khả năng trích xuất dữ liệu từ nguồn dữ liệu thô. Đó chính là công cụ Web Scraper.

Ví dụ đối với việc trích xuất dữ liệu của công cụ Web Scraper:

Dưới đây là một hóa đơn bán hàng online đối với mặt hàng bánh mì nướng nhãn hiệu Warburton thu thập được tại một kho dữ liệu của một siêu thị trong dự án. Nội dung của hóa đơn bao gồm cả dữ liệu có cấu trúc và dữ liệu phi cấu trúc như: dữ liệu hình ảnh (logo nhãn hàng), dữ liệu kí tự (tên sản phẩm, tên nhãn hàng, khẩu hiệu bán hàng…), dữ liệu dạng số (giá sản phẩm, lượng mua, thời hạn…). Thông qua công cụ phần mềm Web Scraper, tất cả các dữ liệu trên hóa đơn sẽ được mã hóa lại thành các dữ liệu có cấu trúc. Tuy nhiên trong khối dữ liệu mã hóa chỉ có một số ít các dữ liệu phù hợp với mục đích tính CPI (như tên sản phẩm, giá sản phẩm). Công cụ Web scraper sẽ tiếp tục trích xuất dữ liệu này để cung cấp cho người dùng tin. (xem hình dưới đây)

Hình 1: Hình ảnh minh họa việc trích xuất dữ liệu thu thập nhờ công cụ Web Scraper

Hóa đơn bán hàng của sản phẩm bánh mì nướng Warburton

Các thông tin trên hóa đơn được mã hóa thành dạng dữ liệu có cấu trúc

```
.....
</div></div><div class="productLists" id="endFacets-1"><ul class="cf
products line"><li id="p-254942348-3" class=" first"><div
class="desc"><h3 class="inBasketInfoContainer"><a id="h-254942348"
href="/groceries/Product/Details/?id=254942348"
class="si_pl_254942348-title"><span class="image"><img
src="http://img.tesco.com/Groceries/pi/121/5010044000121\IDShot_90x
90.jpg" alt="" /><!----></span>Warburtons Toastie Sliced White
Bread 800G</a></h3><p class="limitedLife"><a
href="http://www.tesco.com/groceries/zones/default.aspx?name=qualit
y-and-freshness">Delivering the freshest food to your door- Find
out more &gt;</a></p><div class="descContent"><!----><div
class="promo"><a
href="/groceries/SpecialOffers/SpecialOfferDetail/Default.aspx?prom
oId=A31234788" title="All products available for this offer"
id="flyout-254942348-promo-A31234788--pos"
class="promoFlyout"><span class="promoImgBox"><img
src="/</a></li></ul></div></div></div></div><div
class="quantity"><div class="content addToBasket"><p
class="price"><span class="linePrice">£1.45<!----></span><span
class="linePriceAbbr">(£0.18/100g)</span></p><h4 class="hide">Add
to basket</h4><form method="post" id="fMultisearch-254942348"
.....
```

Kết quả:

Dự án ONS đã tiến hành lắp đặt và vận hành công cụ Web Scraper để phục vụ cho việc thu thập dữ liệu. Mỗi ngày hệ thống phần mềm thu thập được 6.500 các thông tin chi tiết của sản phẩm trong 3 siêu thị bán hàng online, thuộc 35 mặt hàng.

Bước đầu, quá trình thu thập sẽ tiến hành liên tục trong vòng 3 tháng. Một hệ thống quản lý chất lượng sẽ được thiết lập nhằm kiểm soát chất lượng các thông tin có được nhờ việc trích xuất dữ liệu đã thu thập xem liệu chúng có đảm bảo hay không. Sau đó quá trình thu thập dữ liệu sẽ được tiến hành liên tục trong suốt khoảng thời gian tiếp theo của dự án. Kết quả thông tin thu thập được lưu trữ tại cơ sở dữ liệu phi cấu trúc NoSQL.

Bước 3: Phân tích dữ liệu

Các dữ liệu có ích thu thập được sau quá trình trích xuất dữ liệu bởi phần mềm Web Scraper được tiến hành phân tích như dữ liệu có cấu trúc nhờ các công cụ phân tích thống kê như: SPSS, STATA, R, EVIEWS…

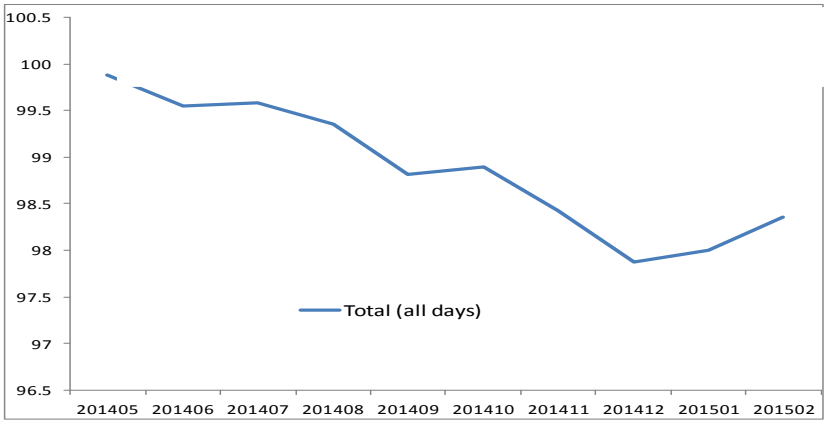
Dưới đây là một số kết quả thu được:

- Tháng 10 và tháng 11 năm 2014 ước tính có khoảng 23% mặt hàng giảm giá, trong đó một nửa sự giảm giá của các mặt hàng được dự tính là giảm giá do mua nhiều sản phẩm (khuyến mại). Như vậy việc giảm giá do mua nhiều sản phẩm là khá phổ biến.

• Sự giảm giá của các mặt hàng có thể được hiểu một cách chi tiết thông qua việc phân tích các dữ liệu trích xuất từ phần mềm Web Scraper. Điều này không thể làm được với các dữ liệu được thu thập bằng phương pháp truyền thống. Nguyên nhân do sự giảm giá thông thường được tính trong trong chỉ số, nhưng giảm giá khuyến mại do mua nhiều thì lại không. Tuy nhiên, dữ liệu trích xuất từ phần mềm Web Scraper chưa giúp người dùng tin nắm bắt được việc giảm giá có nguyên nhân từ chính khách hàng (sự mặc cả giá). Vì vậy kết quả phân tích này cũng chưa phản ánh được toàn bộ vấn đề giảm giá của hàng hóa.

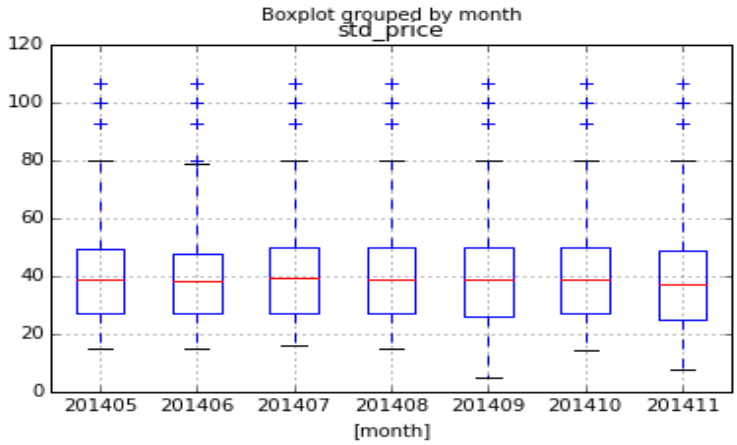
• Sự phân bổ của mô hình giá cả kép và đa mô hình (bi-modal and multi-modal price) thường xuất hiện giữa các mục phân loại CPI/RPI.

Biểu đồ 1: Chỉ số giá Jevon của 35 mặt hàng
thu thập từ tháng 5/2014 đến tháng 2/2015



Nguồn: Big data ONS project - Progress report: Qtr 4 Octorber to Dec 2014 (ONS)

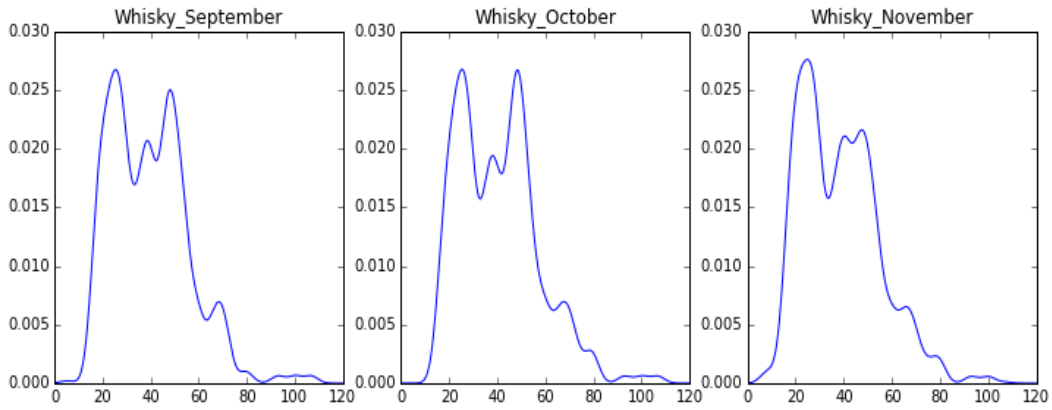
Biểu đồ 2: Biểu đồ hộp ria mèo về giá của 35 mặt hàng
thu thập từ tháng 5/2014 đến tháng 11/2014



Nguồn: Big data ONS project - Progress report: Qtr 4 Octorber to Dec 2014 (ONS)

Một ví dụ khác về việc vận dụng các phần mềm thống kê như SPSS, R…, tiến hành phân tích dữ liệu do công cụ Web Scraper trích xuất ra đối với mặt hàng rượu Whisky:

Biểu đồ 3: Biểu đồ biến động giá rượu Whisky trong tháng 9, 10,11 năm 2014



Nguồn: Bigdata ONS project - Progress report: Qtr 4 Octorber to Dec 2014 (ONS)

Như vậy dự án đã bắt đầu đưa ra được các chỉ tiêu cơ sở có sử dụng dữ liệu Web Scraper. Các dữ liệu thu thập được là các dữ liệu có dung lượng và tần suất lớn cho phép xác định rõ các chỉ số. Điều này không thể thực hiện được bằng những cách thu thập thông thường.

Công việc tiếp theo của dự án sẽ được triển khai theo hướng nghiên cứu chuyên sâu về các vấn đề xoay quanh việc khai thác và phân tích nguồn dữ liệu được trích xuất từ công cụ Web scraper, phục vụ công tác tính chỉ số CPI. Đó là:

- Nâng cao chất lượng Web Scraper
- Quá trình làm sạch dữ liệu (các kỹ thuật, phương pháp lọc, làm sạch thông tin thu thập)
- Nghiên cứu phương pháp phân tích dữ liệu (liệu các phương pháp truyền thống có thể áp dụng cho việc phân tích nguồn dữ liệu mới này hay phải tìm ra một phương pháp phân tích nào khác?)

II. Khai thác nguồn dữ liệu quét[19] từ máy quét mã vạch sản phẩm để tính chỉ số giá CPI tại Pháp

[19] *Dữ liệu quét hay chính là dữ liệu ghi nhớ từ máy quét mã vạch cũng là một trong những loại dữ liệu đặc trưng của thương mại điện tử, chúng không ngừng gia tăng cả về tốc độ, dung lượng lẫn loại hình. Về bản chất thì dữ liệu quét chính là dữ liệu lớn Big data mà chúng ta đang tìm kiếm và khai thác.*

Hiện nay, chỉ số giá tiêu dùng của Pháp cũng đang được tính theo chỉ số giá tổng hợp Laspeyres dựa trên dữ liệu giá cả của các mặt hàng trong rổ hàng hóa đại diện, được các thống kê viên thu thập hàng tháng. Tuy nhiên với việc gia tăng không ngừng số lượng các sản phẩm bán lẻ cũng như những loại hình tiêu dùng đặc biệt (hình thức bán hàng giảm giá, bán sản phẩm có hàng tặng kèm…) khiến cho việc duy trì mẫu mặt hàng đại diện là điều khó khăn.

Bất cứ khi nào bạn mua một sản phẩm trong siêu thị, kết quả thanh toán gói hàng tiêu dùng của bạn cũng sẽ được máy quét mã vạch sản phẩm lưu lại thông qua việc quét mã vạch. Trên mỗi mã vạch có một bộ số quốc tế để phân biệt với nhau. Mỗi ghi nhớ này bao gồm giá trị, lượng bán, nhãn hàng… và mã quốc tế (EAN là số mã của hàng hóa châu Âu). Những thông tin này có thể giúp ích cho thống kê trong việc tính toán các chỉ số. Như vậy cơ sở dữ liệu thống kê về chỉ số giá cũng có liên quan phần nào đến dữ liệu giá.

Chính bởi những lý do này, năm 2009 INSEE (Viện Thống kê và nghiên cứu Kinh tế quốc gia Pháp) đã tiến hành một dự án khai thác dữ liệu bán hàng thu thập từ máy quét mã vạch (gọi là dữ liệu quét). Tháng 9 năm 2012, các chuyên gia của dự án về công nghệ thông tin do INSEE tiến hành đã có thể truy cập đều đặn hàng ngày vào các nguồn dữ liệu bán hàng lưu trữ nhờ máy quét mã vạch tại các chuỗi siêu thị.

Thông qua việc sử dụng dữ liệu quét, dự án sẽ nghiên cứu khả năng của nguồn dữ liệu này trong việc: Tăng được kích thước rổ hàng hóa đại diện và chất lượng các chỉ số giá hàng tháng; Chọn mẫu ngẫu nhiên không chệch đối với các mặt hàng trong rổ hàng đại diện; Ước lượng được chính xác các chỉ số giá.

Quy trình thực hiện dự án được tiến hành theo các bước cụ thể như sau:

➢ *Bước 1: Xác định loại dữ liệu khai thác và chọn mẫu*

Căn cứ vào mục tiêu cụ thể của dự án (khai thác nguồn dữ liệu từ máy quét mã vạch sản phẩm để tính chỉ số giá CPI), dữ liệu mà dự án cần thu thập chính là các dữ liệu liên quan tới giá của sản phẩm hàng hóa như giá cả, lượng bán...

Danh mục các mặt hàng điều tra là các mặt hàng gia dụng thuộc rổ hàng hóa tính CPI. Dưới đây là danh mục 8 mặt hàng được INSEE lựa chọn tiến hành thu thập dữ liệu trong quá trình thực hiện dự án:

Bảng 1: Danh mục và số lượng mặt hàng EAN[20] thu thập

Mặt hàng	Số lượng EAN trung bình đối với mỗi siêu thị
Cà phê	186.3
Dầu ăn	66.3
Gạo	74.9
Yoghurt	224.1
Trứng	24.5
Sô cô la	201.5
Nước ép trái cây	151.6
Phô mai	121.2
Tổng	1050.4

Nguồn: would scanner data improve the French CPI? – INSEE, Consumer Price Statistics Division

Mẫu dữ liệu quét bao gồm dữ liệu về giá và lượng bán hàng tuần của các siêu thị trong suốt 3 năm (2007, 2008, 2009) đối với tất cả các mặt hàng gia dụng. Tổng số các mặt hàng khoảng 1050 mục, thuộc 8 loại mặt hàng. Mẫu dữ liệu quét gồm khoảng 130.000.000 quan sát, xấp xỉ mức tổng hợp dữ liệu lớn nhất của hệ thống phần mềm thông thường.

Việc chọn mẫu các siêu thị thu thập thông tin cũng dựa trên việc xem xét nhiều tiêu chí như doanh thu, thị phần của siêu thị, khả năng truy cập thông tin, thương hiệu… Bắt đầu quá trình nghiên cứu, INSEE chọn ra chuỗi 6 siêu thị lớn (chiếm khoảng 30% thị phần trên thị trường) cho phép các chuyên gia có thể truy cập vào nguồn dữ liệu quét của máy quét mã vạch sản phẩm ở các siêu thị này.

Số mặt hàng được rút ra ở mỗi hệ thống siêu thị tỷ lệ thuận với thị phần của hệ thống siêu thị đó trên thị trường. Để làm được điều này, ta giả định rằng diễn biến giá cả của sản phẩm có thể phản ánh phần lớn việc đàm phán giữa người bán (các chuỗi siêu thị) và người sản xuất (được xác định thông qua thương hiệu trên sản phẩm).

Đơn vị cơ bản của rổ hàng hóa theo phương pháp thu thập truyền thống bao gồm có loại mặt hàng và cửa hàng, còn đối với phương pháp thu thập dữ liệu bằng máy quét mã vạch, thì đơn vị cơ bản là sự kết hợp giữa mã EAN với cửa hàng (chẳng hạn như chai Cocacola có mã EAN là A tại cửa hàng B).

[20] *EAN (European Article Number): hệ thống mã số hàng hóa châu Âu*

Trên đây là những nguyên tắc chọn mẫu trong quá trình thực hiện dự án. Tuy nhiên trên thực tế, quá trình chọn mẫu gặp không ít những khó khăn như:

Doanh thu bán hàng của các chuỗi siêu thị là không đồng nhất. Chẳng hạn với riêng một mặt hàng sô cô la, tiến hành kiểm tra tất cả 1388 mã EAN khác nhau của mặt hàng này, ta thấy chỉ cần 100 mã EAN đã tập trung tới 56% doanh thu bán hàng. Mặt khác, nếu xét toàn bộ các loại giá trong kho dữ liệu thu thập từ máy quét mã sản phẩm của siêu thị, thì rổ hàng lựa chọn có thể bao gồm khoảng 30.000 đơn vị cơ bản. Số lượng đơn vị cơ bản là rất lớn. Vậy đối với những mặt hàng không thể thu thập đủ dữ liệu thỏa mãn yêu cầu của một đơn vị cơ bản thì cần phải giải quyết như thế nào? Điều này dẫn tới 2 ý tưởng.

Một là: việc lựa chọn chuỗi siêu thị có doanh thu bán hàng hàng năm phải tương đương nhau.

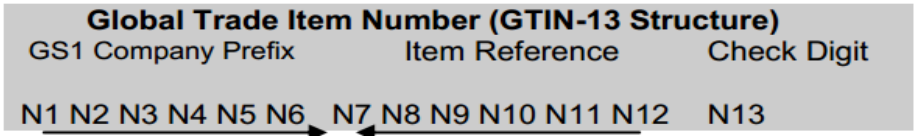
Hai là: hàng tháng các đơn vị cơ bản khuyết thuyết (biến missing) sẽ được thay thế bởi các đơn vị cơ bản khác "gần" với đơn vị khuyết thiếu (đơn vị cơ bản thay thế phải có nhiều điểm tương đồng với đơn vị cơ bản cũ).

Để làm được điều này, tỷ lệ đơn vị cơ bản được chọn thay thế đơn vị cơ bản tương đồng mỗi năm sẽ rơi vào khoảng từ 45-18% số đơn vị cơ bản, tùy theo dữ liệu kiểm tra (thực tế, mẫu kiểm định đã chỉ ra 45% đơn vị cơ bản có trong tháng 12 năm 2008 đã biến mất ở thời điểm tháng 12 năm 2009. Như vậy 13.500.000 đơn vị cơ bản thay thế sẽ phải được lựa chọn). Trong khi đó tỷ lệ thay thế các sản phẩm tương đồng của điều tra viên theo phương pháp thu thập giá truyền thống là 17%. Điều này có nghĩa là tỷ lệ thay thế sản phẩm tương đồng giữa dữ liệu của điều tra viên với dữ liệu thu thập từ máy quét mã vạch hàng năm đối với rổ hàng cố định là khác nhau. Mức độ khác biệt ít nhất là 11%.

Một phần của sự khác biệt là do các chương trình khuyến mại, giảm giá (chẳng hạn như mua 3 tặng 1). Nguyên nhân chính là việc dữ liệu giảm giá vẫn được lưu lại trong máy quét, trong khi với hình thức thu thập truyền thống thì điều này khó để nhận biết. Ngoài ra với hình thức thu thập truyền thống, người hỏi thường chỉ tập trung vào mặt hàng phổ biến, mặt hàng được mua nhiều trong khi đó dữ liệu quét lưu trữ tất cả các dữ liệu, bao gồm cả dữ liệu của những sản phẩm ít phổ biến, ít được mua một cách chi tiết.

Vì EAN là một phần trong đơn vị cơ bản, nên để chọn được một đơn vị cơ bản thay thế cho đơn vị cơ bản bị mất thì ta cũng phải chọn ra được mã EAN tương đồng thay thế cho mã EAN cũ. Cơ sở của việc lựa chọn thay thế như sau:

Mã EAN chính là phần đầu tiên trong mã vạch sản phẩm. Cấu trúc mã vạch sản phẩm như hình dưới đây:



Nguồn: Would scanner data improve the French CPI? – INSEE, Consumer Price Statistics Division

Từ mã vạch đó ta có thể tìm ra được sản phẩm thay thế từ các nhãn hàng tương đồng ở cửa hàng tương tự. Khi một mã EAN trong cửa hàng này biến mất, ta tìm mã EAN đó tại cửa hàng tương tự khác. Nhờ đó ta có thể lựa chọn sản phẩm thay thế với 11 chữ số đầu tiên trên mã vạch giống nhau, nếu không, có thể là 10 số hoặc 9 số… Nếu ngay cả mặt hàng có 6 số mã vạch tương tự cũng không chọn được thì ta tiến hành gán sự biến động giá trung bình của cửa hàng đó cho sản phẩm gia dụng này.

➢ *Bước 2: Phân tích dữ liệu*

Dữ liệu quét thu thập từ máy quét mã vạch là một loại dữ liệu đã được mã hóa. Kho dữ liệu quét của các siêu thị bao gồm hàng loạt các thông tin liên quan đến sản phẩm. Chuyên gia dự án sẽ truy cập vào nguồn dữ liệu này để lấy ra những thông tin cần thiết phục vụ cho việc tính toán CPI. Quá trình phân tích và tính toán được thực hiện bằng các phương pháp và công cụ thống kê phù hợp.

Mô phỏng đầu tiên là về chỉ số lạm phát năm 2009 đối với từng sản phẩm gia dụng được thu thập và với các kích thước khác nhau của rổ hàng hóa (tỷ lệ mẫu chọn là 1%, 2% và 5% tổng số các mặt hàng), 500 mẫu độc lập được rút ra.

Kết quả này cho phép ta đánh giá được độ chính xác của các chỉ số giá được tính từ dữ liệu quét một cách chi tiết. Cụ thể là:

Đối với mẫu tỷ lệ 2%, ta có thể thấy độ dài của 95% khoảng tin cậy của các chỉ số giá, nhỏ hơn độ dài 95% khoảng tin cậy các chỉ số giá của mẫu tỷ lệ 1% (điều này có nghĩa là nếu biến động chỉ số giá trung bình của 500 mẫu tỷ lệ chọn là 3%, thì độ dài của 95% khoảng tin cậy 500 chỉ tiêu tính được từ 500 mẫu loại này sẽ dài hơn 95% khoảng tin cậy của 500 chỉ tiêu với mẫu tỷ lệ 2% và ngắn hơn 95% khoảng tin cậy của 500 chỉ tiêu với mẫu tỷ lệ 4%).

Dưới đây là kết quả tính toán phân tích cụ thể tỷ lệ lạm phát giá trung bình 2009. Kết quả nằm trong một phần báo cáo của Vụ Thống kê giá của INSEE khi thực hiện dự án khai thác dữ liệu quét để cải thiện chỉ số giá CPI ở Pháp:

Tỷ lệ lạm phát trung bình năm 2009 với một số mặt hàng

Kết quả tính tỷ lệ lạm phát trung bình năm với 2 mặt hàng: mặt hàng quan trọng nhất (yohurt) và mặt hàng ít quan trọng nhất trong danh mục 8 mặt hàng được chọn (gạo).

Bảng 2: Tỷ lệ lạm phát trung bình của gạo năm 2009

Tỷ lệ mẫu (%)	Số lượng quan sát rút ra	Tỷ lệ lạm phát trung bình năm 2009 (%)	STD (%)	Min (%)	Q1 (%)	Q5 (%)	Q95 (%)	Q99 (%)	Max (%)
1	350	-2.1	0.58	-3.4	-3.4	-3.0	-1.1	-0.6	-0.3
2	700	-2.1	0.40	-2.9	-2.9	-2.7	-1.4	-1.2	-1.0
5	1750	-2.1	0.23	-2.8	-2.6	-2.4	-1.7	-1.5	-1.3

Kết quả cho thấy tỷ lệ lạm phát năm 2009 đối với gạo là -2.1% và với tỷ lệ mẫu 2% thì 98% của 500 chỉ tiêu không khác quá 1% so với tỷ lệ lạm phát năm trung bình (Q1= -2.9% và Q99= -1.2%)

Bảng 3: Tỷ lệ lạm phát trung bình của yohurt năm 2009

Tỷ lệ mẫu (%)	Số lượng quan sát rút ra	Tỷ lệ lạm phát trung bình năm 2009 (%)	STD (%)	Min (%)	Q1 (%)	Q5 (%)	Q95 (%)	Q99 (%)	Max (%)
1	1795	-4.4	0.23	-5.6	-5.0	-4.8	-4.0	-3.9	-3.7
2	3590	-4.4	0.16	-4.8	-4.8	-4.7	-4.1	-4.0	-3.9
5	8980	-4.4	0.10	-4.7	-4.7	-4.6	-4.2	-4.2	-4.1

Kết quả cho thấy tỷ lệ lạm phát năm 2009 đối với Yohurt là -4.4% và với tỷ lệ mẫu 1% thì 98% của 500 chỉ tiêu không khác quá 1% so với tỷ lệ lạm phát năm trung bình (Q1= -5% và Q99=-3.9%)

Bảng 4: So sánh chỉ số giá các sản phẩm
của dữ liệu quét với dữ liệu điều tra viên thu thập

Tỷ lệ lạm phát 2009	Sôcôla thanh (%)	Nước ép (%)	Dầu ăn (%)	Cà phê (%)	Trứng (%)	Gạo (%)	Yohurt (%)	Phô mai (%)
Chỉ số CPI chung	+0.2	+2.6	-5.3	+2.4	-0.7	-4.0	-4	-3
Chỉ số CPI từ dữ liệu điều tra viên thu thập	-0.8	+2.1	-4.7	+2.5	-1.7	-4.3	-4.3	-2.8
Chỉ số CPI tính từ dữ liệu quét	-0.1	+1.7	-5.9	+2.1	-1.0	-4.4	-4.4	-2.4

Ở đây chúng ta đã so sánh chỉ số giá các sản phẩm gia dụng tính bởi dữ liệu quét với tỷ lệ lạm phát 2009 (toàn bộ dữ liệu về tỷ lệ lạm phát CPI bao gồm tất cả các loại cửa hàng), và chỉ số CPI siêu thị do những điều tra viên tính theo phương pháp truyền thống. Ta có thể thấy chỉ số tính từ dữ liệu quét khá gần với chỉ số CPI siêu thị, sự khác biệt không quá 1% cho tất cả các sản phẩm.

Tỷ lệ lạm phát năm 2009 đối với cả 8 sản phẩm

Chúng ta tính tỷ lệ lạm phát năm 2009 cho cả 8 mặt hàng gia dụng, dữ liệu kiểm tra do tất cả các siêu thị cung cấp, dựa trên các chỉ số dữ liệu quét. Quyền số của các sản phẩm gia dụng được tính thông qua doanh số bán năm 2008 với dữ liệu quét.

Ta cũng có thể ước lượng thông qua các mô phỏng mẫu trong cơ sở dữ liệu quét ở mức 95% khoảng tin cậy đối với các chỉ số giá, tỷ lệ mẫu phụ thuộc vào việc thu thập thực tế của người thu thập. Kết quả thu được như sau:

Mặt hàng tiêu dùng	Quyền số	Chỉ số tính theo dữ liệu quét 2009 (%)	Chỉ số tính theo dữ liệu của điều tra viên (%)	95% khoảng tin cậy với chỉ số tính theo dữ liệu quét 2009	95% khoảng tin cậy với chỉ số tính theo dữ của điều tra viên 2009
Cà phê	15.6	2.1	1.1	0.5	3.7
Sôcôla thanh	11.8	-0.1	1.7	-1.8	1.6
Dầu	8.5	-5.9	-5.1	-8.2	-3.6
Gạo	3.8	-2.1	1.3	-5.8	1.6
Yohurt	21.1	-5.4	-5.7	-5.9	-2.9
Phô mai	15.6	-2.4	-3.6	-3.7	-1.1
Trứng	9.9	-1.0	-2.6	-2.8	0.8
Nước ép	13.6	1.7	0.2	0.2	3.2
Tổng 8 mặt hàng	*100.0*	*-1.4*	*-2.0*	*-2.0*	*-1.1*

Như vậy ta có thể thấy mức độ lạm phát năm 2009 của cả 8 sản phẩm ước lượng từ dữ liệu quét (-1.4%) xấp xỉ với ước lượng mức lạm phát 8 sản phẩm bởi ước lượng từ dữ liệu của điều tra viên (-2%).

Tỷ lệ lạm phát 2009 với tất cả hệ thống siêu thị

Tính các chỉ số giá tiêu dùng 2009 đối với tất cả các chuỗi siêu thị. Sau đó so sánh các chỉ số giá tính được chỉ dựa vào nguồn dữ liệu của điều tra viên với chỉ số kết hợp của cả dữ liệu quét do người bán cung cấp và dữ liệu thu thập của điều tra viên.

Quyền số dựa trên doanh số bán hàng 2008 lấy từ tài khoản quốc gia và từ mẫu dữ liệu quét. Kết quả như sau:

Mặt hàng tiêu dùng	Quyền số	Chỉ số tính hoàn toàn theo dữ liệu của điều tra viên (%)	Chỉ số tính theo dữ liệu kết hợp giữa dữ liệu quét và dữ liệu điều tra viên (%)
Cà phê	27.4	2.5	2.8
Sôcôla thanh	29.7	-0.8	-1.4
Dầu	26.6	-4.7	-4.9
Gạo	28.4	-2.4	-2.0
Yohurts	24.4	-4.0	-3.9
Phô mai	27.5	-2.8	-2.5
Trứng	28.7	-1.7	-1.3
Nước ép	28.9	2.1	2.4
Tổng 8 mặt hàng	*27.3*	*-1.5*	*-1.3*

Ta có thể thấy chỉ số kết hợp (mixed index) của toàn bộ 8 sản phẩm (-1.3%) khá gần với chỉ số dữ liệu hoàn toàn do thu thập (-1.5%)

Những công việc tiếp theo của dự án

Với những kết quả mà dự án khai thác dữ liệu quét đối với lĩnh vực giá do INSEE thực hiện, có thể kết luận rằng dữ liệu quét thực sự là một nguồn dữ liệu đầy tiềm năng trong việc cải tiến chất lượng các chỉ số giá. Cho đến nay dự án vẫn được tiếp tục triển theo các hướng nghiên cứu sau:

- Nghiên cứu sâu hơn về kích thước phù hợp với rổ hàng hóa đại diện
- Thiết kế mẫu đối với rổ hàng hóa đại diện (nghiên cứu kỹ về tỷ lệ tương ứng giữa mẫu rút ra với doanh số bán hàng)
- Cách thay thế các mặt hàng khuyết thiếu hoặc bị mất dựa trên các đặc điểm tương đồng của sản phẩm (liệu có thể quá trình thay thế này có thể xây dựng thành một quy trình tự động?)
- Nghiên cứu sâu hơn về chất lượng các chỉ số giá tiêu dùng được tính từ dữ liệu quét
- So sánh cụ thể giữa hai loại chỉ số giá tiêu dùng: một loại được tính từ dữ liệu quét, một loại được tính từ dữ liệu thu thập bởi thống kê viên.

• Liệu có thể chuyển đổi các phương pháp phân tích dữ liệu truyền thống sang một dạng phân tích phù hợp có thể áp dụng đối với nguồn dữ liệu mới thu thập được không.

Kết luận

Trên đây là kết quả ứng dụng thực tế trong việc thu thập, khai thác và sử dụng nguồn dữ liệu Big data trong công tác thống kê giá. Rõ ràng Big data là một nguồn dữ liệu đầy tiềm năng và đem lại nhiều giá trị thiết thực. Nếu Việt Nam nắm bắt được điều này thì chắc chắn Big data sẽ mở ra rất nhiều cơ hội tốt cho Việt Nam nói chung và ngành Thống kê nói riêng. Trước mắt, có thể ứng dụng vào lĩnh vực điều tra giá tiêu dùng ở một số điểm lấy giá là các siêu thị thuộc khu vực thành phố lớn. Một số điểm cần lưu ý khi áp dụng là: Thực hiện trao đổi, thỏa thuận quyền thu thập, truy cập thông tin đối với các đơn vị cung cấp; Thông tin thu được có thể là dạng có cấu trúc hoặc phi cấu trúc (hình ảnh, âm thanh…) nên cần được sự hỗ trợ của công nghệ thông tin trong quá trình xử lý và làm sạch; Hiệu quả chi phí; Quản lý dữ liệu, hạ tầng Công nghệ thông tin…

Hy vọng rằng trong tương lai không xa, Việt Nam nói chung và Tổng cục Thống kê nói riêng sẽ nắm bắt và khai thác thành công nguồn dữ liệu này.

Tài liệu tham khảo:

1. ONS Big data Project-Progress report: Qtr 4 October to Dec 2014

Jane Naynor, Nigel Swier, Susan William, Karegass, Rob Breton Offical for National Statistics

2. Would scanner data improve the French CPI? –INSEE, Sesbastien FAIVRE-Consumer Price Statistics Division

3. http://searchdatamanagement.techtarget.com/essentialguide/Big-data-applications-Real-world-strategies-for-managing-big-data

4. https://www.acquia.com/examples-big-data-projects

5. http://www.informationweek.com/big-data/big-data-analytics/8-reasons-big-data-projects-fail/a/d-id/1297842

6.https://vi.wikipedia.org/wiki/H%E1%BB%99i_%C4%91%E1%BB%93ng_m%C3%A3_s%E1%BA%A3n_ph%E1%BA%A9m_th%E1%BB%91ng_nh%E1%BA%A5t_ch%C3%A2u_%C3%82u